DEVELOPMENT OF A VISUALIZATION AND INFORMATION MANAGEMENT PLATFORM IN TRANSLATIONAL BIOMEDICAL INFORMATICS

A Dissertation
Presented to
The Academic Faculty

by

Todd Hamilton Stokes

In Partial Fulfillment
of the Requirements for the Degree
Doctorate of Philosophy in Bioengineering in the
School of Electrical and Computer Engineering, College of Engineering

Georgia Institute of Technology
May 2009

DEVELOPMENT OF A VISUALIZATION AND INFORMATION MANAGEMENT PLATFORM IN TRANSLATIONAL BIOMEDICAL INFORMATICS

Approved by:

Dr. May D. Wang, Advisor
School of Biomedical Engineering
Georgia Institute of Technology

Dr. Shuming Nie
School of Biomedical Engineering
Georgia Institute of Technology

Dr. Robert Butera
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. John Petros
School of Medicine, Urology
Emory University

Dr. Alfred Merrill
School of Biology
Georgia Institute of Technology

Date Approved: April 1, 2009

To all of the proofreaders in my life.

ACKNOWLEDGEMENTS

I'm tremendously indebted to my advisor, Dr. May Wang, and to my committee, many of whom I worked with and learned from throughout my training. Dr. Shuming Nie deserves special thanks as the co-Director of the Georgia Tech – Emory Center for Cancer Nanotechnology Excellence (CCNE) which supported much of this work. His pioneering research on molecular imaging using nanoparticles has drawn an impressive group of scientists to the CCNE and enabled the opportunity for me to participate in caBIG and see real impacts of my work. Dr. Robert Butera mapped out his scientific career in a lab meeting one day, and this convinced me that it was not only possible, but very important for researchers with Electrical Engineering backgrounds to get involved in interdisciplinary Bioengineering. Dr. Al Merrill shared with me many details of the workings of his proteomics and lipidomics lab. His interest in understanding the impact of computational technologies goes beyond most biological scientists and led to many insights for integrating software with their work. Dr. John Petros and his lab opened my eyes to the world of health care information security and the unique problems of doing research in a patient care environment. Each of these is a critical piece to the puzzle of information technology translation. Finally, May Wang led my training with vision and foresight into the future of this field. It was a very lucky day when I first walked into her office to discuss her lab and her research interests and found that they could expand my own personal interests in new directions. Dr. Wang embodies the idea that hard work, rigorous training, relentless investigation of the problems of biomedical research, and a passion for understanding will open doors to tackle important problems. I am blessed to have received training under someone of her caliber.

My fellow researchers in the Biomedical Informatics and Bioimaging Laboratory (BioMIBLab) are contributors in various ways to my work. Richard A. Moffitt deserves special thanks for many deep intellectual discussions, preparation sessions, and general

friendship. John H. Phan also helped me tremendously in technical discussions. Besides being a leader for our lab in taking on the challenge of defending his work, he has great skill in resolving issues that impede computational research. C.F. Quo and Jeff Wang provided valuable moral support during my early years as a graduate student. Sovandy Hang, James Torrance, Henry Li, Martin Ahrens, Kathy Pham, Pierre LaRochelle and Randy Han were excellent undergraduate mentees. Our lab offers a unique training experience for biomedical engineers, and they showed great courage and will by accepting ever greater challenges and presenting their work in a rigorous environment.

I also wish to thank organizers Dr. Marie Thursby, Dr. Carolyn Davis, and Kathleen Kurre and fellow students participating in the National Science Foundation Technology Innovation: Generating Economic Results (TI:GER) Fellowship program. The experience of writing a business plan with my team was a valuable extension to my doctoral education. In addition to the interdisciplinary work among science and engineering disciplines, the opportunity to work on a large project with Management & Business Administration and Jurisprudence Doctorate students helped frame the impact of my work in a commercial environment.

I wish to thank my family, my friends, and all of my social supporters. My mom and dad have been great role models and inspiration for my life. My mom was especially significant as an inspiration in pursuing her own Ph.D even after serving a full career as a teacher. I thank Benjamin J. Fierman, Royal P. Madison and Vania E. Stokes for keeping me grounded and sane outside of school.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AJAX	Asynchronous Javascript and XML
ANOVA	Analysis of Variance
APNG	Animated Portable Network Graphics
ASCII	American Standard Code for Information Interchange
caBIG	Cancer Biomedical Informatics Grid
caDSR	Cancer Data Standards Repository
CDE	Common Data Element
DAP	Data Analysis Protocol
DAVID	Databse for Annotation, Visualization, and Integrated Discovery
EVS	Enterprise Vocabulary Services
FDA	Food and Drug Administration
GenMAPP	Gene Map Annotator and Pathway Profiler
GEO	Gene Expression Omnibus
GO	Gene Ontology
GPS	Genome-Phenome Superhighway
GSEA	Gene Set Enrichment Analysis
GUI	Graphical User Interface
GWAS	Genome-Wide Association Studies
HTML	Hypertext Markup Language
IDE	Integrated Development Environment
IND	Investigational New Drug
KNN	K Nearest Neighbors
MAQC	Microarray Quality Control Consortium

NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NDA	New Drug Application
PLIER	Probe Logarithmic Intensity Error Estimation
PNG	Portable Network Graphics
SAM	Significance Analysis of Microarrays
SERS	Suface Enhanced Raman Spectroscopy
SIW	Semantic Integration Workbench
SNP	Single Nucleotide Polymorphism
SVG	Scalable Vector Graphics
TAXY	Theta Alpha X Y (Regression)
TBMI	Translational Biomedical Informatics
UML	Unified Modeling Langugage
UMLS	Unified Medical Language System
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
XML	Extensible Markup Language

SUMMARY

Translational Biomedical Informatics (TBMI) is an emerging discipline expanding beyond traditional bioinformatics, with a focus on developing computational technologies for real-world biomedical practice. The goal of my Ph.D. research is to address a few key challenges in TBMI, including: (1) the high quality and reproducibility required by medical applications when processing high throughput data, (2) the need for knowledge management solutions that allow molecular data to be handled and evaluated by researchers, regulators, and doctors collectively, (3) the need for near real-time, efficient access to decision-oriented visualizations of integrated data and data processing results, and (4) the need for an integrated solution that can evolve as medical consensus evolves, without requiring retraining, overhaul or replacement. The problem statement of this dissertation is that new molecular data is too bulky, clumsy, error-prone, heterogeneous, and difficult to interpret to be useful in a real-word clinic with existing informatics technologies. This dissertation addresses this problem with three overall objectives:

Information Management: To develop novel data handling systems that improve quality, search-ability, and efficient archival and maintenance of high throughput data.

Visualization: To develop a web-based visualization system that enables fast and effective biomedical decision-making using standard formats to represent heterogeneous molecular data.

Translation: To enable standardization and clinical workflow integration of biomedical informatics by contributing to quality standards consortia for molecular data

and deploying semantically annotated solutions to the wider cancer community through cancer Biomedical Informatics Grid (caBIG).

The accomplishment of these aims was guided by 3 design principles. Each of these principles is an important component of robust medical software. The first principle is use of open-source software platforms instead of proprietary formats, because this allows for the investigation of technical problems at the deepest possible level. The second principle is the use of web-based "software-as-a-service" architecture for maximum portability and accessibility of the solution. The third principle is the use of open standards for data formats whenever they exist to aid in interoperability. The choice of technologies such as PHP, MySQL, SVG, and PNG were all guided by these principles. These principles are aligned with those of the caBIG community.

This dissertation resulted in the development and adoption of concrete web-based application deliverables in regular use by bioinformaticians, clinicians, biologists and nanotechnologists. These include: the Chip Artifact Correction (caCORRECT) web site and grid services, the ArrayWiki community microarray repository, and the SimpleVisGrid visualization grid services (including eGOMiner, nanoDRIVE, PathwayVis and SphingoVisGrid).

CHAPTER 1

INTRODUCTION

This introduction will present the four grand challenges to Translational Biomedical Informatics: stability and security, information integration, multi-scale analysis and queries, and translation to clinical workflows. No lone researcher can make a significant impact on these grand challenges, but I present how these problems shaped the development of the Specific Aims of this dissertation.

Motivation for Translation Biomedical Informatics

Bioinformatics has traditionally been concerned with computational molecular biology (e.g. sequence alignment, structure prediction, and molecular dynamics modeling). Translational Biomedical Informatics (TBMI) is an emerging field that focuses on developing computational technologies for real-world biomedical practice. These technologies do not replace the medical professional, but rather empower him/her to achieve higher diagnostic accuracy and to more efficiently treat disease. The goal of this research is to address three key challenges for TBMI enabling bioinformatics to be translated into medicine and positively impact human health. These challenges are directly related to society-wide problems of the cost of health care and the prevention of medical errors that can carry a high societal cost.

The key challenges that differentiate bioinformatics and TBMI are: (1) the high quality and reproducibility required by medical applications when processing high throughput data, (2) the need for knowledge management solutions that allow molecular data to be handled and evaluated by researchers, regulators, and doctors collectively, (3) the need for near real-time, efficient access to decision-oriented visualizations of integrated data and data processing results, and (4) the need for an integrated solution that

can evolve as medical consensus evolves, without requiring retraining, overhaul or replacement.

Some of these challenges have been raised in the field of medical informatics. However, medical informatics remains grounded in the concept of the *patient record*. This has prevented standardization as systems ultimately submit to the immediate needs of treating patients. Over time, meta-data becomes as free-form, individually-tailored, and difficult to mine as the patient whose medical record it represents. TBMI must address these challenges while also building a bridge between these operational and the scientific approaches, opening the floodgates for the vast quantity of data soon to be available without causing the health care system to drown in it.

Figure 1 illustrates the problems with integrating new types of high-throughput molecular data into the patient record by giving examples of the diversity of data that is becoming available and the complexity of the relationships between these virtual representations of the extremely complex state of the patient's overall health.

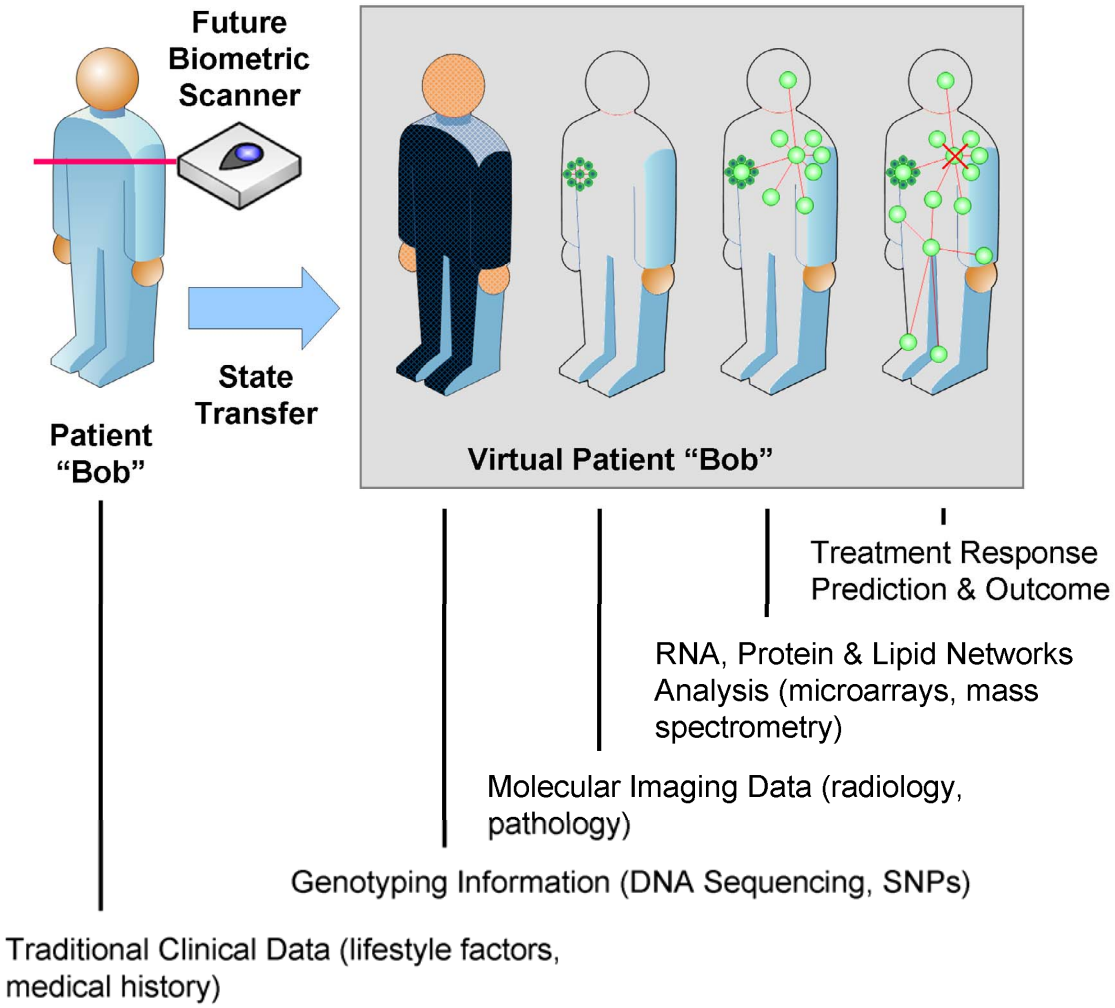


Figure 1: Data management challenges of personalized medicine. As molecular detection technology advances, traditional medical informatics approaches designed to handle only traditional clinical data will become a bottleneck for clinical translation and ultimately for diagnosis in the clinic. Four new types of data are growing rapidly. Genotyping information helps the doctor predict how a patient's systems will react to disease and to treatment based on hard-coded predispositions in his/her genome. Molecular Imaging data gives location and morphology information to further understand the scale of a disease, plan treatment, and evaluate success. Molecular abundance data forms a large-scale biochemical profile which can be compared to a database of previous patients and outcomes to make predictions about disease aggression or treatment response. Finally, all of these predictions must be updated, aggregated and compared to other patients to supply more knowledge for the benefit of future patients.

The Three Grand Challenges of Biomedical Informatics

Information technology has been much slower to impact the medical community than other communities in research and in industry. There are very good reasons for this delay. First, there are the problems of the expense of acquiring data and the resulting data scarcity because the health care research environment is heavily regulated in the interests of societal ethics. Second, there are the further ethical issues around personal data privacy and de-identification once it has been acquired. Finally, there are issues of countless complicating external factors in medicine and health that make quantitative analysis and interpretation of results extremely challenging. All of these contribute to the problem statement of this dissertation: that new molecular data is too bulky, clumsy, error-prone, immature in acquisition methods, heterogeneous, and difficult to interpret to be useful in a real-word clinic with existing informatics technologies.

Stability and Security

The innate ethical sensitivity of medical data means the need for security is of primary importance. The time-sensitive nature of delivering medical care means that stability is also critical. Both of these challenges are related in that an unstable system is more vulnerable to security attacks. Many high-profile data leaks in recent years have demonstrated that human factors (e.g. not following data security policies or installing systems that have not been reviewed) are a major source of security holes. However, in the defense of human intelligence, data security policies are too complicated and most software systems do not use simple techniques such as gentle reminders or automated security audits to help users prevent intruders.

Two important aspects of the stability and security problem that guided design principles of this dissertation are the use of data formats that may contain corrupt, invalid data (or even malware developed by cyber attackers) and the use of proprietary systems that may not receive regular security updates if the development team is not committed to

the ongoing support of the system. Data that can be validated by human and by automated computer programs is an important advancement for improving stability and security of medical software. Another important advance is the advent of community-based and non-profit computing platforms like The Wikipedia Foundation and the NCI Cancer Biomedical Informatics Grid (caBIG), which do have a public charter to create long-lasting and secure resources for the common good.

Information Integration

TBMI is an interdisciplinary field with very few participants that have the good luck to be trained and practiced in all aspects. For this reason, in any meeting of biologists, biomedical engineers, computational scientists, and clinicians, some audience members can be confused by discussions that go to a substantial amount of technical depth (whether they admit it or not). This is due to the importance of language training in integrating new information into the human brain. The more comfortable someone seems in using a word that has no meaning for the listener, the more uncomfortable the listener gets and the harder it becomes to effectively communicate.

Turning to computational technologies can help overcome these barriers. Resources like Wikipedia (http://www.wikipedia.org) allow for deep investigation into new subjects at a much faster rate than when reference books had to be located, browsed and digested. However, for computational algorithms to accomplish the same feat, they require more than simple text found in web articles and scientific literature to differentiate between similar concepts. Ontologies are an important tool for allowing computer software to make these distinctions with the same ease as the human brain. Semantic annotation of data is the process of mapping all data using identifiers from a common ontology so that computers can reference the meaning of data by performing a database look-up.

The best example of a medical research community tackling the problem of information integration is the National Cancer Institute's (NCI) Cancer Biomedical Informatics Grid (caBIG) [1]. This national initiative has been held up by the Director of the National Institutes of Health (NIH) as an example for other medical fields to follow. Already, new projects such as the Cardiovascular Research Grid (http://www.cvrgrid.org/) from the National Heart, Lung, and Blood Institute are being proposed and are entering their pilot phases based on the framework and success of caBIG.

Advances in eCommerce and eScience have shown that better information technology infrastructure can stimulate long-term scalability, community involvement and synergy. However, biomedical infrastructure projects face unique challenges such as (1) the rigorous scientific validation demanded by the ethics of medical practice, (2) the urgency arising from deploying systems into an environment designed for daily confrontations with human illness, and (3) the organizational problems that naturally arise in multidisciplinary collaborative research. Addressing these challenges requires more than just the technology approach, but the solutions presented in this dissertation will be limited to how information technology enhances the process of translating to the clinic [2, 3].

Multi-Scale Analysis, Queries, and Data Summaries

Biomedical informatics visualization research can be traced back to the 1970s, when Game of Life [4], a two-dimensional cellular automaton invented by John H. Conway was published in Scientific American. For many years, until the completion of the Human Genome Project in 2001 [5-7], bioinformatics tools were primarily (1) simulations with limited underlying experimental data (high-throughput methods such as microarrays and mass spectrometry were not commonly used); and (2) comparative tools

with limited test cases (the fundamental similarity of the genetic code for all of life had not been confirmed) [8].

In the post-genome era, advanced high-throughput biotechnologies generated large amounts of experimental data, and the birth of Internet during the past decade has made much of these data accessible for effective worldwide collaboration. This has spurred the bioinformatics community to search for innovative techniques to manage and explore these data to drive new discoveries [9, 10]. Many predict that this flood of data is just beginning, which makes translational bioinformatics an exciting field for modern biomedical research [11]. For example, with much more data there is a possibility to model the flow of information in living systems for us to get a deeper understanding of how these systems are sustained and what mechanisms cause them to break down.

Biomedical researchers today are overwhelmed by the quantity of data, and are underwhelmed by the limited usability offered by many tools. This has caused the generation of knowledge, which requires human intervention, to lag behind the generation of data. Quality integrated systems will speed up new discoveries in biology and medicine by allowing researchers to increase the scale and scope of their investigations. Knowledge can be generated by large-scale integrated systems at a rate similar to experimental data as technologies for biological data handling and comparison become more sophisticated. With each successive large-scale project publication, interactive utilities for exploration and visualization of data become more important. Examples of innovative interactive visualization and exploration tools include Databse for Annotation, Visualization, and Integrated Discovery (DAVID) [12], GeneWindow [13] and the Genome-Phenome Superhighway (GPS) [14, 15]. In addition to accessing a vast data resource of value for primary research, good visualization technologies are also portable and simple to access. For visualization in the clinical setting, standard applications should be used to launch visualizations without requiring support from the local information technology experts.

The Organization of This Thesis

This dissertation addresses the challenges mentioned above by applying engineering design toward three fundamental software metrics: interoperability, usability, and adaptability. The three overall objectives of this research were: (1) Information Management: To develop novel data handling systems that improve quality, search-ability, and efficient archival and maintenance of high throughput data, (2) Visualization: To develop a web-based visualization system that enables fast and effective biomedical decision-making using standard formats to represent heterogeneous molecular data, and (3) Translation: To enable standardization and clinical workflow integration of biomedical informatics by contributing to quality standards consortia for molecular data and deploying semantically annotated solutions to the wider cancer community through cancer Biomedical Informatics Grid (caBIG).

These objectives lay out a technical framework for facilitating necessary steps in clinical translation. Chapters 2 and 3 of this dissertation address two components of the Information Management solution: quality control and data discovery. Chapter 4 addresses the Visualization solution. Chapters 5 and 6 address two components of the Translation solution: the Food and Drug Administration's (FDA) Microarray Quality Control (MAQC) Phase II effort and integration of the concrete deliverables into caBIG. Each chapter covers background and significance, gives a technical explanation of the system design, and presents results and documentation of the concrete deliverable. Figure 2 shows how the tools discussed here map to the engineering design factors and the chapter organization. Many tools represent a combination of two factors, but are presented in the chapter that is most relevant. It is expected that this research platform will be used to support ongoing translational research, both at Georgia Tech and Emory and in the wider community.

Information Management

The first objective of this dissertation was to develop novel data handling systems that improve quality, search-ability, and efficient archival and maintenance of high throughput data. This aim can be measured by how well a technology improves interoperability of data. The key metrics of interoperability for TBMI are data quality, portability of data formats, and public sharing of data.

Quality Control

Information management enables information to be passed between many participants in an information workflow. An example of this is web lab experimenters passing data acquisition results to computational experts for analysis. The trust between collaborators in this field is founded on good data quality control. My key contribution in this aim is the development of a web-based system for evaluating the quality of high-throughput microarrays. The quality is evaluated in two ways: visual heatmaps of the data variance that can help laboratory experimenters diagnose problems in their protocols and an algorithmic quality score that can estimate the relative quality of high-throughput experiments by scoring each individual chip. Some have suggested that understanding the timeline that an experiment was performed on is one of the most important indicators of quality [16]. We agree with this assessment and have emphasized tracking of data acquisition time to highest possible level of detail.

Information Integration and Data Discovery

The second component of the first objective of this dissertation was the development of a community-maintained platform for microarray data using a novel data compression and visualization format. ArrayWiki supports information integration and data discovery because of the union of a generalized information representation syntax that is both human- and machine-readable, pioneered by the Wikipedia Foundation. This

platform, ArrayWiki, is at the center of the axes in Figure 2 because it incorporates all of the aims of this dissertation.

ArrayWiki facilitates translation because of the Wiki framework, which involves all community members in discussions of data quality and annotation of the multitudes of microarray experiments, including many clinical experiments. This community involvement is a method to ensure that data remains fresh and up-to-date. ArrayWiki also represents an advance in visualization because it uses the BioPNG format for data compression and storage (discussed in Chapter 4). ArrayWiki supports the effort for better quality control because over 20,000 microarray chips have been imported and automatically processed by the caCORRECT algorithms. The cleaned chip data is available for download by any user on the Internet. Finally, ArrayWiki is most relevant for information management, because the Wiki framework can be extended to support many other types of molecular data, such as mass spectrometry, molecular imaging, or Surface Enhanced Raman Spectroscopy (SERS) nanoparticle data.

Web-based Visualization for Effective Interpretation and Integration of Data

Biomedical Informatics has experienced an explosion in data analysis tools in response to the dramatic growth of molecular data. The biomedical researcher from a non-computational background is similarly bewildered by this array of tools as they are by the task of data management. The second objective of this dissertation is to develop a web-based visualization system that enables fast and effective biomedical decision-making using standard formats to represent heterogeneous molecular data.

This aim addresses challenges of enabling efficient access to information in a fast-paced environment. This aim can be measured by how well the technology improves usability of tools. The key metrics of usability for TBMI are accessibility through the web, flexibility in visual rendering, and responsiveness even when handling high-throughput data.

The visualization system described here is designed to be available on a national computational grid called caGrid, which was built by the caBIG initiative. The system supports visualization of biochemical networks, semantic (or ontological) relationships such as Gene Ontology, correlations and correspondences between analysis results such as mining for biomarkers in molecular data, and high-throughput data acquisition results such as molecular imaging and microarray scanners. The key contribution of this research is that it expands existing technology standards to allow for embedding of the complete source data for a visual representation into the same file that provides the visual data.

Translation

At an information crossroads between many interdisciplinary fields (e.g. molecular biology, chemistry, engineering, and medicine), TBMI researchers face a challenge to develop solutions that facilitate increasingly complex research workflows (see Figure 3). The third objective of this dissertation is to enable standardization and clinical workflow integration of biomedical informatics by contributing to quality standards consortia for molecular data and deploying semantically annotated solutions to the wider cancer community through cancer Biomedical Informatics Grid (caBIG).

This aim can be measured by how easily the wider informatics community can incorporate new analytical processes into existing data workflows. This ability is related to a software engineering design factor called adaptability.

FDA MAQC Phase II Consortium Case Study

The second phase of the Food and Drug Administration (FDA) Microarray Quality Control (MAQC) project studied common methods for building models (i.e. biomarker mining and classification algorithms) from microarray data to predict disease outcomes or drug response. Thirteen clinical scenarios (called “endpoints”) were designed using six datasets (a total of 2276 microarray samples). Thirty-six teams trained classifiers using this data and a wide diversity of data analysis protocols (DAPs) over a

period of one year. Validation data was released to these teams after each one had publicly selected their "best model" from all of their research. These models were finally compared using this blind validation and another performance result was obtained by training on the new validation data a second time and testing on the original data. DAPs were evaluated in terms of simplicity, reproducibility (i.e. low performance variance on new data), and reliability (i.e. predictable performance on new data). One classifier found to perform well was K Nearest Neighbors (KNN). The goal of my contribution to this study was to determine to what extent KNN parameters explained the performance difference between KNN DAPs and to what extent extrinsic properties like data quality played a role. This large-scale example of Team Science is important because the FDA is responsible for approving new technologies like microarrays for use in real-world clinical practice.

caBIG Silver-Level Certification

The cancer Biomedical Informatics Grid (caBIG) is a large-scale initiative of the National Cancer Institute (NCI) to develop a common software platform for all of cancer research. This effort is currently moving from the Pilot Phase (where the emphasis has been on building tools and proving the infrastructure) to the Enterprise Phase (where the emphasis is on adoption and training the clinical research community). In the Enterprise Phase, universities and private companies have been encouraged to deploy their tools to the grid to add value to the overall program. A certification system was set up with Bronze, Silver, and Gold-level certification requirements to guide contributing software developers toward high quality interoperable systems. The Gold-level certification is only available to tools that have already passed Silver-level, and the best practices aren't ready for mainstream developers. My contribution to this effort was to build a package for Silver-level certification of caCORRECT Grid Services and to design SimpleVisGrid to be ready for the same process. The certification review process can be thought of as a

peer review process to decide if systems already accepted as science by the community (as evidenced by publication) can be called Team Science.

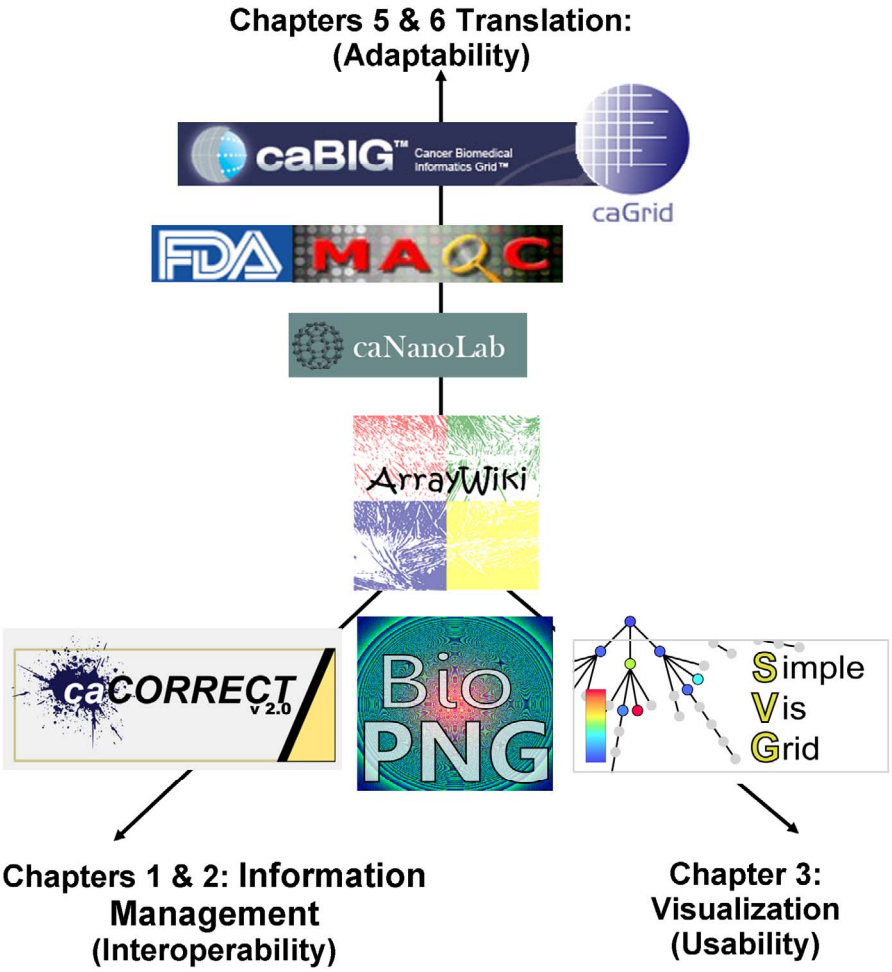


Figure 2: Overview of tools developed (or adopted in the case of caNanoLab and caGrid) for this dissertation. The tools presented here are organized by the most relevant objective or design factor. Chapter 2 will discuss caCORRECT and quality control. Chapter 3 will discuss ArrayWiki and information management. Chapter 4 will discuss BioPNG and SimpleVisGrid. Chapter 5 will discuss the FDA MAQC project (application of quality control to the regulatory community to facilitate translation). Chapter 6 will discuss caBIG Certification efforts for caCORRECT and SimpleVisGrid Silver-level Compatibility Review.

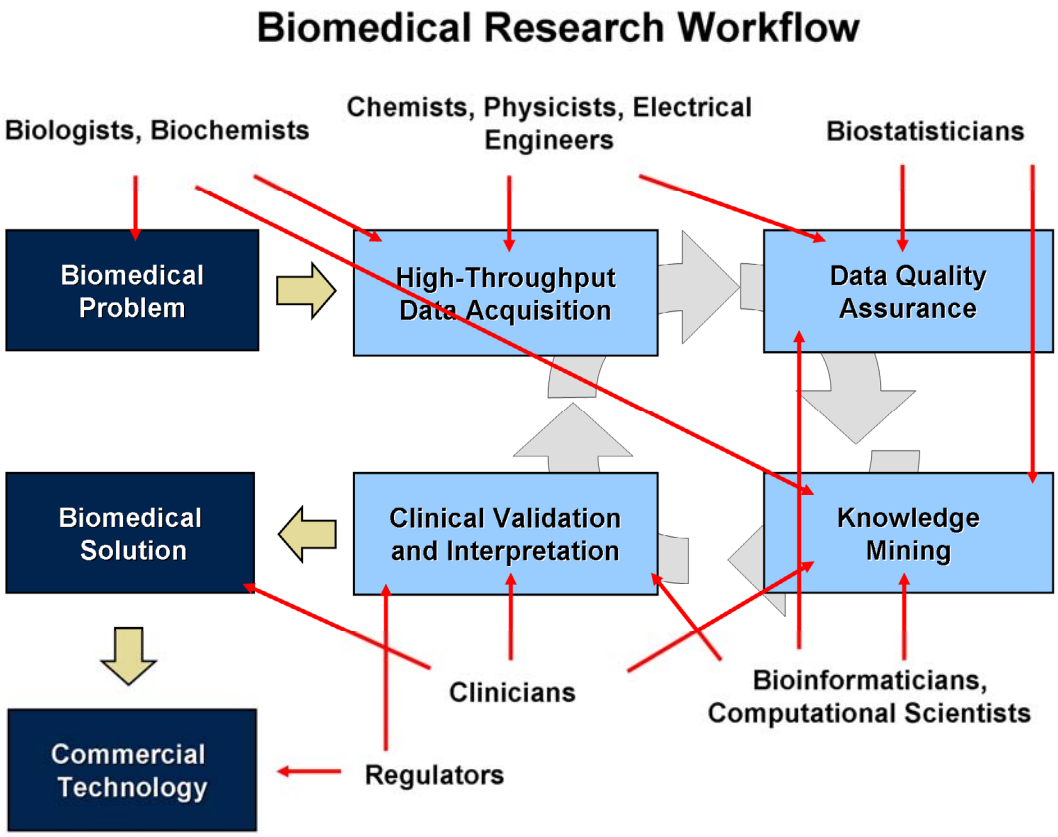
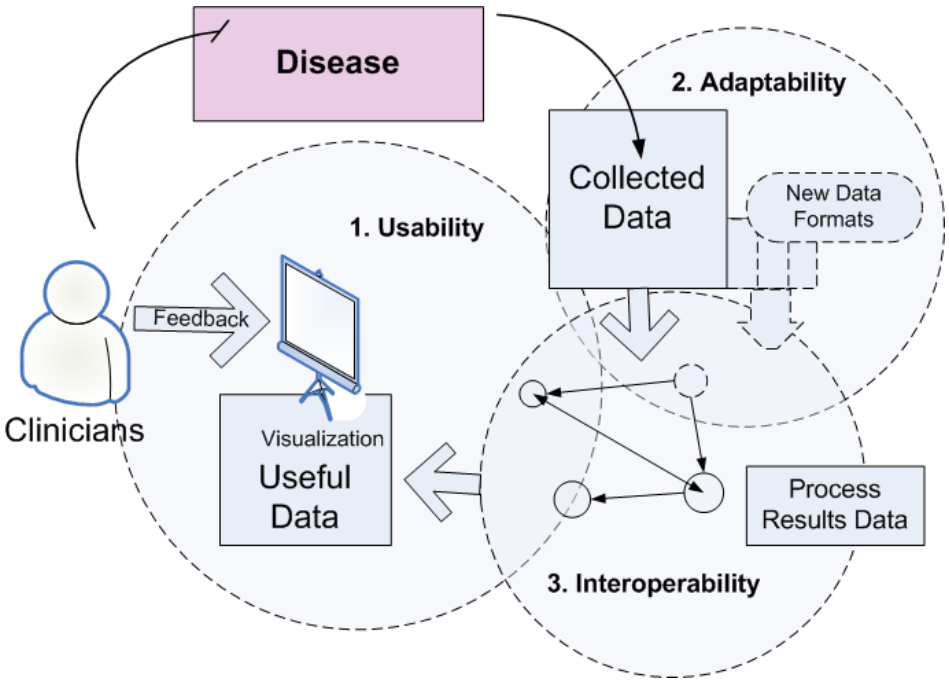


Figure 3: The Biomedical Research Workflow. As interdisciplinary research, Translational Bioinformatics must focus on technologies that are useful and understandable to scientists from many backgrounds. Tools developed specifically for one step in this workflow must interoperate with tools at other steps to avoid friction (i.e. time wasting) and quality problems associated with data transfers and data formatting. For example caCORRECT is specifically designed for "Data Quality Assurance", but it must interoperate with "Knowledge Mining" tools like those studied under the FDA MAQC, and with "Clinical Validation and Interpretation" tools like ArrayWiki because problems validating a biomedical solution may trace back to problems in the "High-Throughput Data Acquisition" step.

The System Design Factors

Three computational design factors will be discussed in this dissertation, because they influence both the choice of technologies and the priorities used in daily tasks of developing robust software. These design factors are usability, adaptability, and interoperability. I have identified the typical approaches to addressing the three computational design factors, and I highlight the importance of standardization for improving all three. In addition, the concrete deliverables of this dissertation represent technical solutions to the stated problems. For interoperability and adaptability, data quality standards and interfaces to a large-scale community project is demonstrated by caCORRECT. For adaptability, ArrayWiki is an example of an even more flexible community-based resource for enabling the community to monitor the currency of data. For usability and interoperability, the grid-based visualization technologies BioPNG and SimpleVisGrid combine open-standard formats and semantically-annotated service descriptions for use by caCORRECT, ArrayWiki, and many other domain-specific applications described in Chapter 4. Figure 4 illustrates the complex relationship between users, systems, and the design factors that make systems more useful for medical research. These developments are of importance not only to bioinformaticians, but also to clinicians and to government agencies such as the National Institutes of Health and National Science Foundation [17].



A.

B.

Design Factors	Usability	Adaptability	Interoperability
Practical Units	Web-based Tools	Knowledge	Process Data
Techniques	Visualization	Wiki, Semantic Annotation	Standards, Ontologies
Scientific Roles	Clinicians	Computer Sci.	Bioengineers

Figure 4: Translational Bioinformatics System Design Factors. **A.** Using software to stop disease. All biomedical researchers are interested in usability and verifiability of the system. The end user should be protected in some ways from the complexity "under the hood." **B.** Most bioinformatics research prototype systems are not friendly enough for use by people without computational knowledge. Computer Scientists have the knowledge of algorithms to extract useful knowledge from data, but they don't have the technical background to acquire high quality data or fully understand the data acquisition process. Bioengineers must understand the data and how to draw relationships between data to connect these two groups and create useful systems.

CHAPTER 2

CHIP ARTIFACT CORRECTION (CACORRECT) QUALITY CONTROL FOR ARRAY DATA

The first objective of this dissertation was to develop novel data handling systems that improve quality, search-ability, and efficient archival and maintenance of high throughput data. This aim can be measured by how well a technology improves interoperability of data. The key metrics of interoperability for TBMI are data quality, portability of data formats, and public sharing of data. This design factor is introduced first in this chapter, and then the background for microarray quality control is presented. Finally, results of caCORRECT are presented, based on its original publication in *Annals of Biomedical Engineering* [18]. Finally, new directions are discussed.

Interoperability

Interoperability is defined as the ability for independently-developed software systems to share information. Interoperability is important because it enables software systems to expand functionality with minimal duplication of development effort. Interoperable modules may be plugged into research workflows to easily test new approaches to data analysis. Interoperability adds a new layer of complexity to integrated systems and requires stable data standards and infrastructure.

For some time in the early days of bioinformatics, true interoperability was avoided by the development of huge data warehousing efforts [19]. Enormous centralized databases were proposed supposedly to contain all of biological knowledge. These efforts were usually soon abandoned due to costs in keeping data current [20, 21]. The underlying reasons for these failures are the distributed nature of biological knowledge into specializations and the dynamic and unpredictable nature of ongoing research and discovery.

Currently, most computational researchers agree that distributed solutions are the most likely to last in such a dynamic field. No organization can possibly implement innovative solutions to every biological problem. Web services have become the vision of the future of e-Science, including bioinformatics [9, 22]. The importance of web services is that those organizations with specialized information to share with the community have freedom to change the underlying data models behind their services without creating instability for other systems that use their data. The only requirement is that their services continue to support the pre-defined transaction types (a.k.a. application programming interfaces (APIs)). This decoupled model between data consumers and developers also improves adaptability. Still, there are many research teams that follow a stand-alone software distribution model where developers must compile and release platform-dependent application versions and support them until they can persuade their users to upgrade locally installed software.

The web services model is ideal for creating integrated solutions because no single organization is required to maintain all of the specialized working parts. Web-based services tend to lead to more stable and robust software because replacing a service can be as simple as changing a single URL, allowing software to evolve by replacing obsolete components quickly as new solutions are developed. One missing component of this model is a set of standards for communicating about the quality of the services that are offered, and thus human intervention is still necessary for systems to evolve.

Adoption of community standards is a recurring theme throughout this dissertation. Figure 5 is designed to show relationships between communication standards developed for computer scientists, those that have been applied to bioinformatics, and the corresponding move from specificity to generality in the data contained in these standards.

Interoperability Standards

The lowest three levels in Figure 5 deal with interoperability standards. In the early days of Internet computing, transport interoperability was the primary data handling technique, including transferring tapes and floppy disks between laboratories. Computer networking and the World Wide Web greatly eased the physical burden of this layer. Some of the greatest recent advances in interoperability are the Transmission Control Protocol (TCP-IP) (1974), Hypertext Transfer Protocol (HTTP) (1990), and Extensible Markup Language (XML) (1996). At the Data Types layer, many developers often wrote custom data formatting code for each platform. XML standards have greatly reduced the need for custom code.

Interoperability in TBMI is more complicated. Teams of collaborators are often composed of experimental scientists who specialize in collecting data and bioinformaticians who specialize in analyzing data. These teams require increasingly advanced methods to ensure that data was collected and transferred with the correct associated meta-data. Data handling and quality control involves proper identification of outliers that were caused by noisy data measurement techniques. The system must employ reliable identification techniques to ensure that it is comparing "apples to apples" as opposed to throwing out important variations that might lead to an important discovery. For this reason, the development and adoption of data identification standards is also critical to the success of this approach.

Data Identification Standards

The vast array of identification schemes for genes, proteins, metabolytes and other biological species adds complexity to the interoperability of data handling systems. Web services such as GeneCruiser [23] have been developed to assist in translation between differing identification schemes. The FDA Microarray analysis workbench, ArrayTrack [24], also contains ID translation functionality, built on top of the Bioconductor [25] package for the R programming environment. Data identification

standards may be the route to improving the ability to merge experiments that contain heterogeneous data acquisition platforms. A good introductory effort to integration of heterogeneous data is the Biozon database [26], which is an attempt to collect data from the most often-cited databases and provide a unified search tool that can categorize and rank results based on the Google PageRank [27] model. The categories available in Biozon are Protein Families, Pathways, Proteins, Domains, Structures, Interactions, Nucleic Acids and Unigene Cluster. The Life Sciences Identifier (LSID) [28] shows the most promise for uniting these identification schemes, but due to the immature status of this standard, it does not play a definite role in interoperability today.

Ontologies: Vocabulary Standards

Vocabulary standards are collectively known as ontologies. The goal of biological ontologies is to precisely define the vocabulary used in research, so as to reduce confusion between researchers that use the same words or glyphs (graphical indicators) to identify or describe very different entities. A common example is the word 'agent', which can indicate a biologically active molecule, an autonomous computer program, a government worker, or a financial representative, etc. depending on the context.

Bioinformatics tools have necessitated the development of ontologies because they are increasingly expected to interpret the language of researchers. This requires the ability to cross-reference terms and place them appropriately in an increasingly complex biological context. The impact of these ontologies, created initially by bioinformaticians, will be felt throughout the medical and biological research communities as the definitions of common words are necessarily narrowed, broadened, or reassigned for clarity in the computational setting.

Ontologies are 1) dictionaries of terms with definitions, usually organized in a conceptual hierarchy, rather than alphabetically; 2) a formal way of organizing discussions around what people mean when they use certain terms; and 3) a conceptual

map of a domain, which may be used for visualization or calculations of distances between ideas.

Ontologies as a systematic specification of language for a large community are not an entirely new concept. Mathematics and electronic circuit theory are two familiar examples of ontologies that include terminology and glyphs. While some might describe the development of these ontologies as a laborious process, it is hard to dispute their success in allowing scientists from different disciplines to communicate specific problems and solutions with minimal confusion.

Biomedical ontologies are expected to have a future impact comparable to the impact of mathematics and electronic circuit theory. They will demystify the biological area of knowledge by breaking down concepts into their component parts, allowing researchers to move between different levels of detail in their discussions while (hopefully) avoiding semantic conflicts that inhibit the discovery process.

Biomedical Data Quality and Data Sharing Standards

Many interdisciplinary groups have published papers about the lack of reproducibility of biological experimental results due to lack of information provided about data acquisition protocols and data analysis protocols [29-31]. Initiatives such as Minimum Information About a Microarray Experiment (MIAME) [32] and Minimum Information Requested in the Annotation of Biochemical Models (MIRIAM) [33] have attempted to set guidelines for meta-data sharing to improve reuse of existing experimental data for new purposes. Finally, quality surveys (including Phase I one of FDA MAQC) have shown that microarrays are susceptible to many sources of noise, and that noise reduces the reproducibility of downstream results [16, 34-37]. These standards efforts in the microarray field have been mirrored in related high-throughput fields such as mass spectrometry [38] due to recognized quality problems [39, 40].

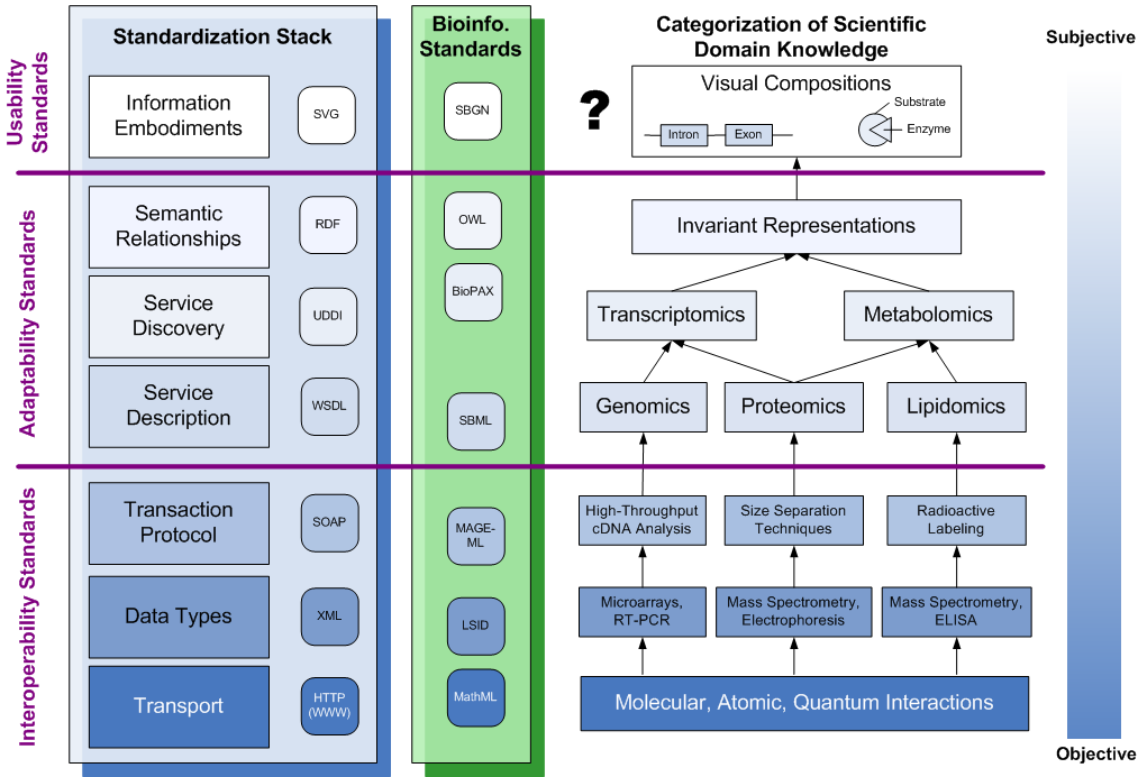


Figure 5: The Translation Biomedical Informatics Software Standardization Stack. This chart presents the progression of developments toward standardization of software system design. It is inspired by and an extension of the Network Communications Protocol Stack (see http://en.wikipedia.org/wiki/Protocol_stack for details). Progressing upwards through the chart, it is important to note that standardization becomes more difficult as requirements move from the Objective (or concrete) type to the Subjective (or personal preference) type. The question mark at the top indicates that usability standards are difficult to identify and represent important future work. The parallel upward progression of knowledge integration in the life sciences field is meant to help non-computational readers to understand the change from concrete requirements, such as syntax or timing requirements for very specific devices, to general requirements such as how to represent over-arching concepts in information embodiments. Just as life sciences have become rooted in the foundation of understanding molecular (and even quantum) interactions, software design is rooted in the foundation of information transport. However, as concrete data gives way to abstract ideas, it becomes difficult to reach consensus, and therefore the standards must become more flexible.

Chip Artifact Correction (caCORRECT) for High-Throughput Quality Control

I developed a web-based tool to support the first objective of this disertation. caCORRECT is a web-based quality assurance process for Affymetrix microarrays. Affymetrix is the most popular and most widely available microarray technology and our system supports any of their 50+ chip platforms. caCORRECT has garnered an international user base and has been deployed to the Cancer Biomedical Informatics Grid (caBIG) in the form of semantically annotated grid services. The caCORRECT system has been run on over 20,000 microarray samples as part of loading the ArrayWiki community web site. Finally, we have used caCORRECT to support the Food and Drug Administration (FDA) Microarray Quality Control (MAQC) Phase II effort to validate the reproducibility of predictive models based on microarray data.

Quality assurance of high throughput "–omics" data is a major concern for biomedical discovery and translational medicine, and is considered a top priority in translational biomedical informatics (TBMI). caCORRECT is a web-based bioinformatics tool for chip artifact detection, analysis, and correction, which removes systematic artifactual noises that are commonly observed in microarray gene expression data (see Figure 6). We designed caCORRECT to have several advanced features: (1) to uncover significant, correctable artifacts that affect reproducibility of experiments using data visualization and image processing techniques; (2) to improve the integrity and quality of public archives by removing artifacts; (3) to provide a universal quality score to aid users in their selection of suitable microarray data for new experiments. All of these features make microarray data more reproducible and thus more interoperable by making data sharing more successful. caCORRECT is freely available for on-line use at: http://cacorrect.bme.gatech.edu. caCORRECT already has many active users worldwide, including at Georgia Tech/Emory University, North Carolina State University, Louisiana State University, University of Hong Kong, and University of Essex.

Importance of Microarray Experiments for Medical Research

One of the earliest applications of gene expression microarray data to human medical studies was an effort to subtype two kinds of leukemia [41]. For microarrays to reach their full potential as a clinical molecular profiling tool for personalized and predictive medicine, the quality of microarray data must be addressed. The FDA started the MicroArray Quality Control (MAQC) consortium and is seeking to develop FDA guidelines on microarray quality control and data analysis [42, 43]. However, the current status of microarray quality control and noise reduction is still a collection of scattered tools and methods. While tools such as dChip [44], RMAExpress [45], Harshlighting [46] and SmudgeMiner [47] include methods to improve the quality of microarray data, these tools fail in several important aspects: (1) they do not provide sufficient visualization to help a novice user understand the source of data problems; (2) they do not incorporate spatial information into the outlier detection methods; (3) they do not incorporate outlier information into their normalization routines; and (4) they do not generate dataset quality metrics to help users select high-quality data [48, 49].

To improve the quality of genomic data, it is important to understand the source of the errors and the current state-of-art in quality control. Recent studies have shown that the choice of microarray platforms is important, but not always the primary factor influencing data quality produced by laboratories. Instead, laboratory techniques are often responsible for the lack of reproducibility in microarray datasets [50]. It has even been suggested that some gene co-expression in microarray chips is the result of spatial artifacts—with the gene pair correlations being more a function of relative chip distance than chromosomal distance [51]. Even worse, the methods that are designed to alleviate such chip to chip non-uniformity could actually hamper results [52, 53]. Using caCORRECT's capability for interactive visualization, we discover several classes of artifacts which can be easily linked to their root causes. Among the most common artifacts are scratches, edge effects, and bubble effects that manifest as visible localized

variations in the microarray dataset. These localized variations are not detected at the level of gene expression, but can be seen using low-level scanner outputs and by preserving the original spatial orientation of the microarray.

Much work has already been done at the gene expression analysis level to detect outlier data points and to improve the reproducibility of microarray results. Affymetrix microarrays, for example, can be processed with Affymetrix's own Microarray Suite (MAS5.0), GeneChip Operating Software (GCOS), or Probe Logarithmic Error Intensity Estimate (PLIER), but alternatives such as dChip [44], RMAExpress [45] or Guanine Cytosine Robust Multi-array Average (GCRMA) implementations in Matlab or the R statistical language (http://www.bioconductor.org) also exist. These programs include good measures such as normalization, background correction, and robust model fitting in an attempt to determine gene expression from multiple probe values. Many of them provide a visualization feature showing where outlier probes, or probe sets are located on the chips, but yet they do not include this spatial information in their outlier detection schemes. Direct comparisons of caCORRECT to these methods are difficult because caCORRECT is a quality assurance step happening before expression analysis, and with each of the methods mentioned above, noise removal and gene expression calculation are inseparable.

One method by Reimers and Weinstein [47] does take spatial effects into account. This system can be used to visualize regional biases across high-density chips. Citing factors such as temperature, liquid flow rate, RNA diffusion rate, and edge effect, they showed that significant regional biases are common and can greatly affect downstream results. In addition to localized background calculation, Reimers and Weinstein's program produces a comprehensive quality score for each chip by measuring the correlation of each probe's expression level to that of its neighbors. While this application may provide quality score information, it does not allow correction of these

artifacts. Users are then faced with a difficult choice to abandon a chip, or to proceed knowing that artifacts exist.

In the development of new methods for quality control and assessment, Brodsky et al. proposed a novel method of using clustering of gene expression profiles across microarrays to indicate quality [54]. First, gene expression profiles are clustered, and then the uniformity of the clusters' distributions across the microarrays are measured. Second, the patterns of high and low expressed genes are monitored on each sample for uniformity. These two methods provide a dual description of a gene's artifactual nature, which is then used to discard it from further analysis. Most of the genes discarded in this way are identified as a result of artifacts localized to one or few chips. Brodsky's strategy of removing such a gene from the experiment entirely is too harsh, and therefore caCORRECT uses the more conservative strategy of removing offending gene data from the specific chips which contain the artifacts. In this way, caCORRECT is able to retain all genes on the array while removing potentially distracting noise.

A brief survey of current microarray databases (e.g. Gene Expression Omnibus (GEO) [55], arrayExpress [56], caArray [57], Center for Information Biology Gene Expression Database (CIBEX) [58, 59], and Standford Microarray Database (SMD) [60]) reveals that quality analysis at this level of detail is beyond the scope of many labs which produce microarrays today. The post processing done by labs corresponds to the goals of their experiment. Some labs produce technical replicates in hopes to increase signal-to-noise ratios and to reduce the need for laborious artifact detection, while others just ignore these effects and look only at large-scale data features. For data curators and data consumers (who often only have access to the published expression data, and not to the more detailed output of the scanner), a different approach is needed. Our goal in designing caCORRECT is to make public data a knowledge resource for the whole community, and let goal-oriented researchers with more detailed goals use data at a level of detail appropriate to their investigations.

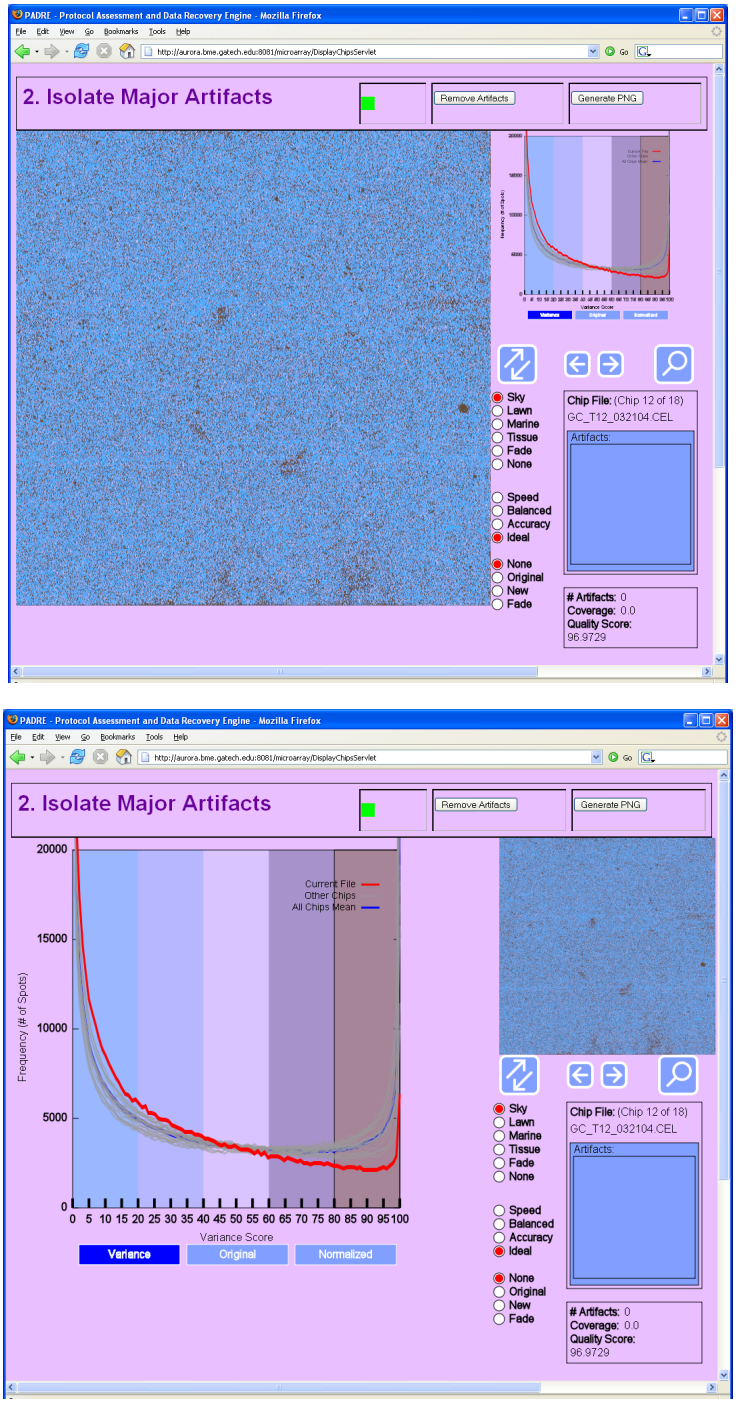


Figure 6: caCORRECT Interactive Mode Screen Shots. These are two examples of the caCORRECT interactive mode interface under normal use. The default configuration is on top. Clicking on the switch button (two opposing arrows) produces the image at the bottom with the enlarged histogram plot. The color stripes in the histogram plot correspond directly to the color bands used in the heatmap. The histogram panel can display the distributions of variance score (shown), original intensity values and normalized intensity values.

System Overview

The caCORRECT workflow (see Figure 7) centers on detection and removal of regions of probes causing localized chip variances (a.k.a. artifacts). The defining feature of an artifact is that it clearly results from errors in microarray manufacturing or lab processing, and not from the underlying biological state being measured. The first step in the caCORRECT workflow is a modified quantile normalization process to align the distributions of each uploaded chip and remove global chip biases. Following this step, variance scores are calculated to analyze data quality on a probe by probe basis. Next, image processing is run on the variance data to identify artifacts. At this point, quality metrics are calculated describing the artifact coverage and noise content of each chip and of the experiment as a whole. Outlier detection is an iterative process because many small to medium artifacts are over-shadowed by larger artifacts in earlier rounds, but become detectable once the data is renormalized using the artifact-aware process.

Upon completion of caCORRECT, the user is presented with the following files: 1) heatmap images of all of the chips, with and without artifact masks, 2) new versions of 'clean' probe expression files with appropriate data replacement, and 3) gene expression value tables calculated by the Bioconductor implementation of PLIER using data before and after caCORRECT. Figure 6 shows an example of caCORRECT's interactive interface, developed to help the user understand how the variance calculation and artifact detection works. Most users prefer the automated batch mode after their first experience with the tool. In batch mode, the entire process of artifact removal runs without input from the user.

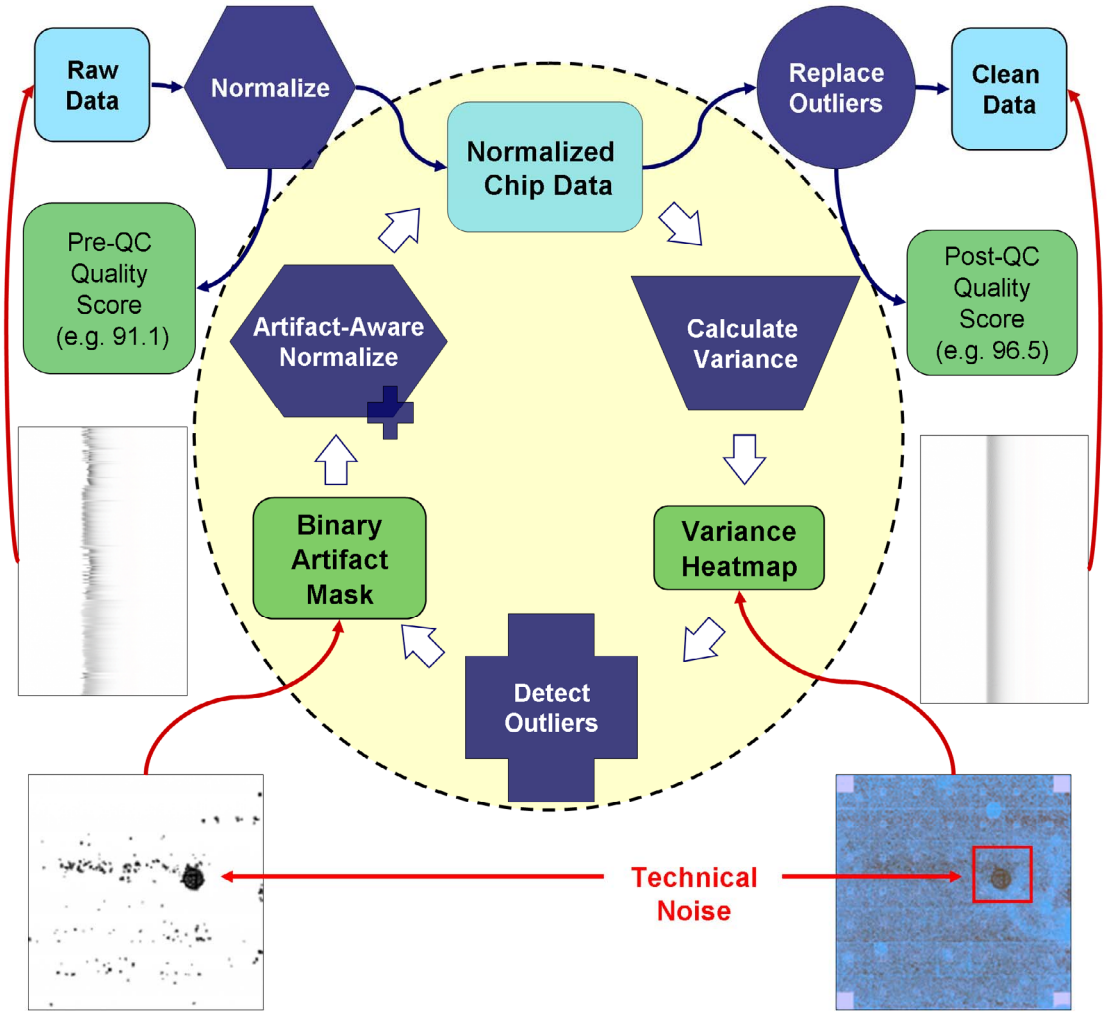


Figure 7: caCORRECT Workflow Diagram. The core workflow components include Quantile Normalization, Variance Calculation, and Artifact Detection. The automated mode runs all steps up to replacement, while the interactive mode allows users to adjust parameters, remove problematic chip files, and update results.

Heatmap Generation

The heatmaps provided by caCORRECT serve to aid the user in validating the results of the system, as well as to see and learn about the nature of the artifacts in the data, and perhaps to take steps (e.g. updating lab protocols) to avoid recreating these mistakes in the future. Heatmaps are created by applying a threshold to the variance score (roughly the 80th percentile for scores.) This threshold may be adjusted in interactive mode, to make the system more or less sensitive to variance. Any variance value above that threshold is automatically assigned to the 'hottest' heatmap color. This color indicates data that will activate artifact detection when it is concentrated to regions on the chip.

For a microarray chip containing reliable data (unadulterated by experimental protocol errors), these hot spots will represent real mRNA concentration differences in the experimental sample. Modern layout techniques for microarrays ensure that these spots will be distributed randomly throughout the chip. In many cases, however, protocols do not achieve uniform hybridization due to uneven drying, formation of salt streaks, scratching or contamination of the microarray surface due to contact with skin or dust, miscalculated hybridization times, or failure to control environmental variables [61]. All of these most common mistakes result in clearly visible localized patches of high variance (a.k.a. artifacts) on the heatmap.

Data Quality Scoring

A major goal of the caCORRECT system was to describe experimental datasets and chips according to comparable quality scores. The desired properties of a quality score for microarray experiments are as follows: (1) If chip variation between technical or biological replicates is low, the quality score is high; (2) the quality score of an experiment composed of identical chips is maximum; (3) The quality score should be bounded for easy comparability (e.g. [0,1]); (4) Masking of high variance regions from

the scoring algorithm should improve quality; and (5) the ordering of the chips should not affect the quality score. Our efforts have produced two balanced scores: the Uniformity Score (described below) and Artifact Coverage Percentage.

For the Uniformity Score, we implemented a pairwise Normalized Cross-Correlation (NCC) algorithm that satisfies all of our desired properties. The score gives feedback to a lab which is generating microarray data as to the quality and repeatability of their own work. It also gives users of public microarray data repositories some criteria by which to select a dataset. In addition to this uniformity score, we calculate an artifact coverage percentage during the artifact identification process for each chip, and for the dataset as a whole. This number can be used to remove highly suspect chips from the data set to improve analysis results. These scores are easily integrated into one overall score by averaging uniformity score with 1-Artifact Coverage and are converted to a scale from 0 to 100 for easier interpretation. It is recommended that any chip with an overall score less than 80 (roughly 25% artifact coverage) be considered for removal from the dataset, especially if the artifacts are widespread and diffuse in nature, as opposed to a sharp, localized artifact that is easily removed.

Artifact Detection and Artifact-Aware Gene Expression Calculations

Perhaps the most obvious benefit of this tool is the identification and replacement of artifact-flagged data before they can foul downstream results. caCORRECT accomplishes this by sending its own heatmaps through a battery of image processing routines which are designed to identify spatially relevant areas of high variance. The image processing routines can vary from relatively simple moving-window searches used in batch mode to complex custom-designed kernels used in interactive mode to find specific artifacts based on user input.

To maintain the specificity of artifact removal during interactive mode, we require that users provide minimal information or parameters regarding the type, size, and

location of probable artifacts present on the chip image via a simple point-and-click interface. Based on the artifact type and size suggested by the user, an image kernel (mask) is generated. Simple morphological convolution is then applied to the chip heatmap image based on the custom kernel. Depending on the type of artifact being identified, various open and close operations are then performed to further specify the artifacts. Once identified, these identified artifacts may be superimposed on the original chip image for visual comparison by the user. Flagged spots already indicated by proprietary software provided by Affymetrix encoded into the CEL file format can also be superimposed on the image for comparison during the artifact flagging process. In most cases, we have found that outliers specified in CEL files do not correspond well to our artifact definitions, but are more randomly distributed throughout the heatmap. The user has the option of retaining these outlier indexes, or replacing them upon data retrieval.

Aside from visual inspection, we suggest data quality metrics (discussed later) as a quantitative measure of the completion of the artifact removal process and resultant chip quality. The user may accept the identified artifacts at any time, or repeat the artifact removal process iteratively until they are satisfied.

Once the user is satisfied that the artifact removal is complete (or at the end of batch removal) the data may be retrieved by the user with the artifact-flagged data appropriately replaced. The most common methods are to replace individual probe intensities with the mean or median values of that probe for each non-artifact sample in the dataset. Besides these options, caCORRECT also allows replacement with zeros, with 'null', or any other custom value that the user prefers.

Quality Assessment of Public Data

Our artifact removal strategies were validated against real data from the scientific literature. We selected recent experiments from ArrayExpress and Oncomine with

available Affymetrix CEL and an associated publication. Our results (see Table 1) show three interesting trends. First, the yeast genome chips produce data of overall higher quality than the human microarrays. These chips also represent the most recently acquired data. Second, in cases where artifact coverage is the greatest, the improvement in quality score due to running caCORRECT is also the greatest. Finally the quality score corresponds well to predicted similarity of the biological samples. Separating chips by class produces the highest quality scores when the classes are the most uniform (wild-type yeast and healthy lung tissue) while samples of low uniformity (modified yeast and tumor samples) are generally of lower quality.

Table 1: caCORRECT quality score results on public data.

Database	Chip Model	Experiment	Initial Quality	Clean Quality	Chip #	Artf. Covg.
Array-Express	YG_S98	Yeast NOS stress – All [62]	95.15	95.817	20	2.35%
		Wild-type only	97.368	97.613	12	3.60%
		Other Types	92.382	94.347	8	4.75%
Oncomine	HU6800	Lung Carcinoma – All [63]	91.577	93.441	98	3.88%
		Tumor Class I	90.393	92.878	67	4.36%
		Tumor Class III	91.187	92.842	19	5.55%
		Healthy Lung	97.711	98.007	10	1.64%
Oncomine	U95Av2	Breast Carcinoma – All [64]	92.076	92.531	89	1.30%
		High-risk	92.231	93.148	18	3.27%
		Low-risk	91.482	92.258	19	1.87%
		Recurring	91.483	91.882	18	1.65%
		Non-recurring	91.799	92.498	34	1.63%

Extension to New Affymetrix Platforms

By design, caCORRECT can support any chip scanned on the Affymetrix platform that produces CEL files in their standard formats. However, to ensure that caCORRECT stays current with the latest chip designs, it was tested with two new types of chips from Affymetrix. One is the HG_HT_U133B, which is designed for running 96 samples in one scan. I encountered these chips while analyzing the Connectivity Map dataset from the Broad Institute at Massachusetts Institute of Technology. These chips are automatically recognized by the software, but interpreting the resulting artifact locations is more difficult because there are many more control regions in the chip.

The second type of chip was provided by the Affymetrix HapMap Project for Genome-Wide Association Studies (GWAS) [65-67]. These new Genome SNP chips are much, much larger than previous chips. They measure up to 1,000,000 single nucleotide polymorphisms (SNPs) and are generally considered to be a more robust use of microarrays than those developed for gene expression. These new chips are rectangular rather than square. Luckily, all of our algorithms were developed to deal with rectangular chips. Table 2 shows the advances in scale presented by these new chip platforms.

Table 2: Relative sizes of new chips supported by caCORRECT.

Dataset Name	Connectivity Map Build 2	Dr. Andrew Young Renal Fixed Tissue	HapMap
Chip Platform	HT_HG_U133B	X3P (biggest processed to date)	GenomeWide SNP v6
Dimensions	744 x 744	1164 x 1164	2680 x 2572
Total Intensity Measurements	553536	1354896	6892960
Percent Increase Over Previous		244%	508%
Number of Experiment Samples	6029	24	270
Typical Quality Score	90	72	75

This rapid change in the scale of data sets for processing was anticipated in the research plan for caCORRECT. However, certain infrastructure improvements and optimizations were required as chips doubled and then quadrupled in size. These improvements include:

1. Migration of the entire code base to a 64-bit test and production server (named Hercules and Gaia). This allows the software to surpass the 4GB memory limit of 32-bit machines. The new memory limit is now 16GB.

2. Conversion of intermediate data files from ASCII text to binary formats. This makes debugging more difficult, so utilities were developed to readily convert between the two formats, and all processes can switch between the two.

3. Transition from the Bioconductor-based PLIER implementation to our own custom gene calculation method (called TAXY, or Theta Alpha X Y Regression).

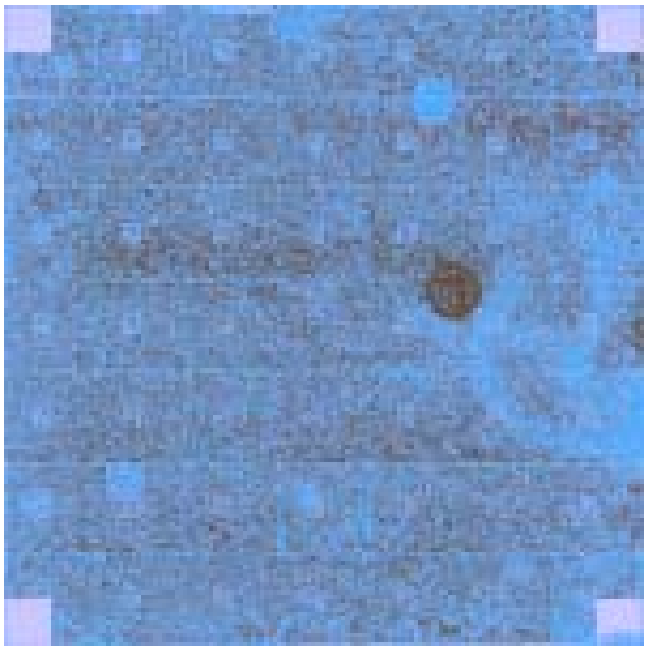
PLIER does not support the exclusion of artifact-flagged probes from the gene calculations. This meant we had to use median replacement of artifact probes as a pre-processing step to the gene calculations. The R implementation of PLIER also did not scale very well to datasets of large sample size, mainly because of memory management limitations. The combination of these two factors made the gene calculations step the largest contributor by far to the run time of our processes. Our newly-developed method is significantly faster (a factor of 10x) and uses memory more efficiently (by loading each

probe set one at a time). This new algorithm is called TAXY and was developed by Richard Moffitt and Weiguang Wang.

Integration with ArrayWiki Community Repository

The full caCORRECT quality control and analysis pipeline is run during the ArrayWiki import process. The variance heatmap and artifact mask images are displayed on the main experiment page. The quality scores are also available in the table of samples on the main experiment page. The original expression calculations can be downloaded from the experiment information box and the "clean" expression calculations are also available. Clean expression data are either median-replaced artifacts run through PLIER or artifact-aware TAXY calculations depending on when the experiment was imported.

(A)



(B)

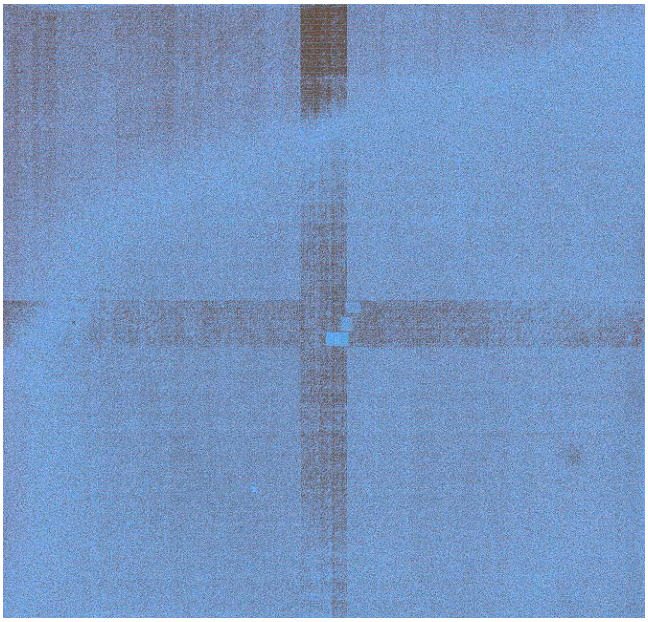


Figure 8: New chip platforms supported by caCORRECT. (A) Example of a caCORRECT variance heatmap for a high-throughput gene array from Affymetrix. (B) Example of a caCORRECT variance heatmap of a HapMap SNP chip with a significant edge effect.

CHAPTER 3

ARRAYWIKI FOR COMMUNITY-BASED INFORMATION MANAGEMENT

This dissertation describes a platform for Translational Biomedical Informatics (TBMI). ArrayWiki is the centerpiece of this platform. Not only does ArrayWiki display the results of running caCORRECT for thousands of experiments, but it also serves as a repository for all microarray-based analyses run in our laboratory. The Wiki framework natively supports the visualization technologies described in Chapter 4, so those features will be associated with their associated experiments. Although the design factor of adaptability maps most clearly to the third objective of translation, it is important to begin the discussion of adaptability here because the flexibility of the Wiki framework is ideal for uniting a research community around a single technical topic. In this way, ArrayWiki is also a step toward translation, though it is too focused on microarrays to ever achieve the impact that caBIG tools will have. This chapter is based on the publication of ArrayWiki in *BMC Bioinformatics* [68].

A survey of microarray repositories reveals that their contents and data models are heterogeneous, and that repositories developed for different communities have become "silo-ed" over time. Bioinformaticians have difficulties in finding datasets based on technical parameters rather than keywords. A community maintained resource, ArrayWiki, resolves these problems by uniting disparate meta-data that are difficult for users to find. This resource provides users with simple text-based searches across all experiment metadata, and exposes data to search engine crawlers (Semantic Agents) such as Google to further enhance data discovery. In addition, automated quality control processes provide extra information about data quality not available in other microarray resources. The data in this resource are open to community contribution, comment, and

modification, and are distributed and visualized using a novel, compact data storage format, BioPNG. ArrayWiki is available at http://www.bio-miblab.org/arraywiki.

Adaptability

Adaptability ensures that when opportunities arise for a service to be used in a different scenario, the architecture will not require major modification to suit the new circumstances. Making each piece of the design modular and using common standards to represent inputs and outputs to each module is the primary way to achieve adaptability. Thus, adaptability and interoperability are fundamentally related. Adaptability may be measured by the useful lifetime of a software product.

Adaptability also includes the ability to upgrade the infrastructure under a service without adversely affecting its availability. Web services allow this flexibility by well-described, robust methods of distributing HTTP requests among clusters of server computers. Individual servers can be taken off-line for hardware upgrades as long as the remaining servers can handle the expected increase in load. Improved up-front system design does not guarantee performance because of the unpredictability of user load in a distributed environment.

Background on Large-Scale Integration Efforts

As shown in Figure 9, current bioinformatics integration projects vary widely in respect to scale and in application of each of the three design factors described above. Based on the failure of unification projects, such as the Integrated Genome Database project [21, 69], most bioinformaticians now agree that distributed solutions are the most likely to succeed on the long term. This has resulted in most projects having the common ground in their selection of basic interoperable technologies. This dissertation presents three well-known bioinformatics integration systems, each mapped to a design factor: myGrid [70] is presented here, Genome-Phenome Superhighway (GPS) OmicBrowse [14] is presented in chapter 4, and caBIG [71, 72] is presented in chapter 5.

MyGRID

Of the integration efforts designed to improve adaptability of bioinformatics tools, BioMOBY and Taverna myGrid [73] represent the grass-roots efforts. All integration projects depend on voluntary participation by the many bioinformatics labs around the world. BioMOBY [74, 75] took the voluntary aspect one step further by allowing service providers to define data types for the inputs and outputs of their services. BioMOBY consists primarily of a central registry of services (called MOBY Central). Each service entry in the registry contains lists of input and output object types, a URL, and a description of the service type. Along with this comes a library of data structure templates (MOBY Objects) and two hierarchies, one for data structures and one for services. Each of these three resources is completely open for modification by the scientific community.

Taverna is workflow-building software linked with the myGrid vision. It is one example of a class of bioinformatics tools centered centered around task composition [76]. The workbench screen makes a default library of web services available. There are also mechanisms for searching for additional web service registries or importing web service descriptions found independently. Web services from the library can be dragged onto a workflow diagram and connected by specifying data interfaces. Using this tool, bioinformatics researchers can experiment with connecting various tools together into bigger and more feature-rich applications. Software developers can explore the value of integrating services and address certain unexpected data interface issues before beginning the user interface design process. Like caBIG, the acceptance by the bioinformatics community of this architecture is evidenced by the efforts of other groups to extend the work into new domains [77].

	GPS OmicBrowse	Taverna myGrid	caBIG
Usability	GPS OmicBrowse uses intuitive visualizations of chromosomes custom to each supported organism	Taverna requires a standalone install. Assembling and troubleshooting workflows requires knowledge of WSDL and XML.	caBIG is still primarily focused on tools for developers. End-user applications still use standard web forms.
Adaptability	GPS OmicBrowse does not interact with other applications.	Taverna allows you to plug in additional services as they become available on the Web. Custom interfaces connect them.	caBIG is focused on the cancer research community and some of its solutions may not translate to others.
Interoperability	GPS OmicBrowse does not import or export data in standard formats.		caBIG collects standards related to every aspect of medical research and publishes them to it's community with best practices.

Figure 9: Summary of Strengths of Existing Large-Scale Integration Projects.

Community-Based Data Maintenance: ArrayWiki

The result of our quality survey of public microarray data is being deposited into an innovative Wiki system, ArrayWiki, for public access and community modification. The creation of ArrayWiki necessitated a novel data format, BioPNG, which adds visualization, portability, and quality information in addition to compression.

Previous surveys of microarray repositories reveal that their contents and data models are heterogeneous, and that contents of repositories developed for different communities have diverged over time. We have identified three usability problems with existing repositories. First, bioinformaticians have difficulties in finding datasets based on technical parameters (e.g. chip scan date) rather than biological keywords because most repositories are queried using a "literature search" paradigm. Second, when performing meta-analyses on microarrays (e.g. merging datasets, re-assigning biological classes, or removing low-quality chips), there is no community resource for storing new results linked in an intuitive way to the source data. Third, when merging data from different sources for comparative studies or to increase statistical significance, the maximum level of detail is necessary to standardize protocols and minimize bias from the separate data sources.

We developed ArrayWiki to address these problems. The data in this resource are open to community contribution, comment, and modification, and are distributed and visualized using a novel, compact data storage format, BioPNG. Domain scientists can make a significant impact in this community by making a small investment of time to learn the syntax and structure common to all sites running MediaWiki software, and contributing to this knowledge base. ArrayWiki is available at http://www.bio-miblab.org/arraywiki.

Comparison to Existing Public Microarray Repositories

ArrayWiki was the result of studying the strengths and weaknesses of existing microarray data repositories (see Figure 10). These include Gene Expression Omnibus (GEO) [55, 78], ArrayExpress [56], caArray Database [79], Stanford Microarray Database (SMD) [49], and oncoMine (OM) [80, 81]). This is only a sampling of the many online gene expression repositories, but GEO and ArrayExpress are the largest repositories by far. More recently, a group from University of California, Los Angeles published Celsius [82], an effort to merge all Affymetrix data from disparate repositories into one location, available through a single programmatic interface. The Celsius authors support the importance of this work for three main reasons: the microarray repository field became very fragmented, data at the CEL file level is difficult to find even in the largest repositories, and experiments are annotated inconsistently across repositories.

All of these databases represent important efforts for ensuring that resources spent on microarray experiments are not lost, but are preserved for future generations of researchers [53]. However, most of these databases fail to provide any chip quality information. Also, they do not offer a familiar Wiki interface for community data curation without using a programmatic interface. Finally, none of these repositories have made a noticeable effort to include the Affymetrix DAT file type in their experiment records. The DAT files available in ArrayWiki offer the highest possible detail level about public experiments and allow bioinformaticians to more deeply explore data quality and improvements on the algorithms used by Affymetrix software (see Figure 11). ArrayWiki is not intended to replace these public repositories of data, but will augment the information they contain with information provided by novel algorithms (caCORRECT) and by the community.

Figure 12 depicts the overlap in experiments between four popular repositories. A standard procedure was used to generate this figure. All datasets examined were public, and had submission (or release) dates between August 2005 and June 2006 inclusive.

Each dataset was searched in every other database using no date criteria. The criteria for determining matching datasets were species, platforms, authors, affiliation and publication (if available). This was repeated for each database. Our interpretation of Figure 3 is that repositories developed for different communities have become "silo-ed" over time. The majority of experiments are found in only one repository (1358+528+10+7=1903 or 80%). Experimenters tend to patronize a particular repository, and the only evidence of an effort to merge repositories with the purpose of facilitating large-scale data mining is the incorporation of SMD experiments into ArrayExpress and GEO at certain points. This means that bioinformatics researchers must search all repositories to ensure they've collected all public data relevant to a particular topic.

Finally, despite the cost of obtaining tissue samples and the complexity of analysis and interpretation, human and mammalian chips still outnumber all other sample organisms (e.g., cell lines, plants, and single-celled organisms). This statistic may be inflated by the failure of many repositories to distinguish between samples taken directly from human tissues and those from genetically modified human cell lines.

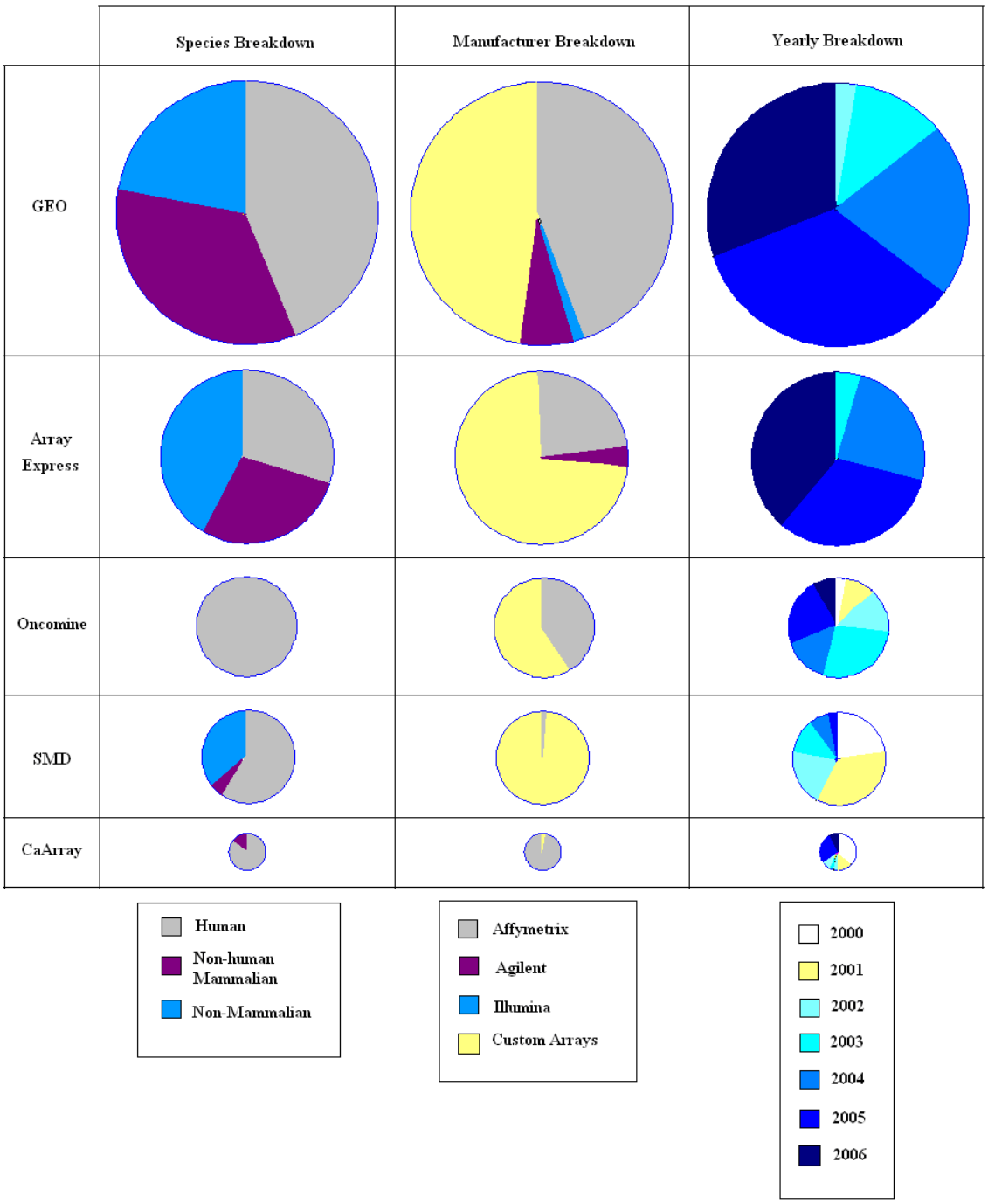


Figure 10: Comparison of microarray repository contents. The relative size of each pie corresponds to the size of each repository. Key observations include that SMD does not contain much recent data. One data artifact is found in the caArray Yearly Breakdown. An abnormal number of experiments show a date of '1-1-2000' because that is the default and the validation is not adequate.

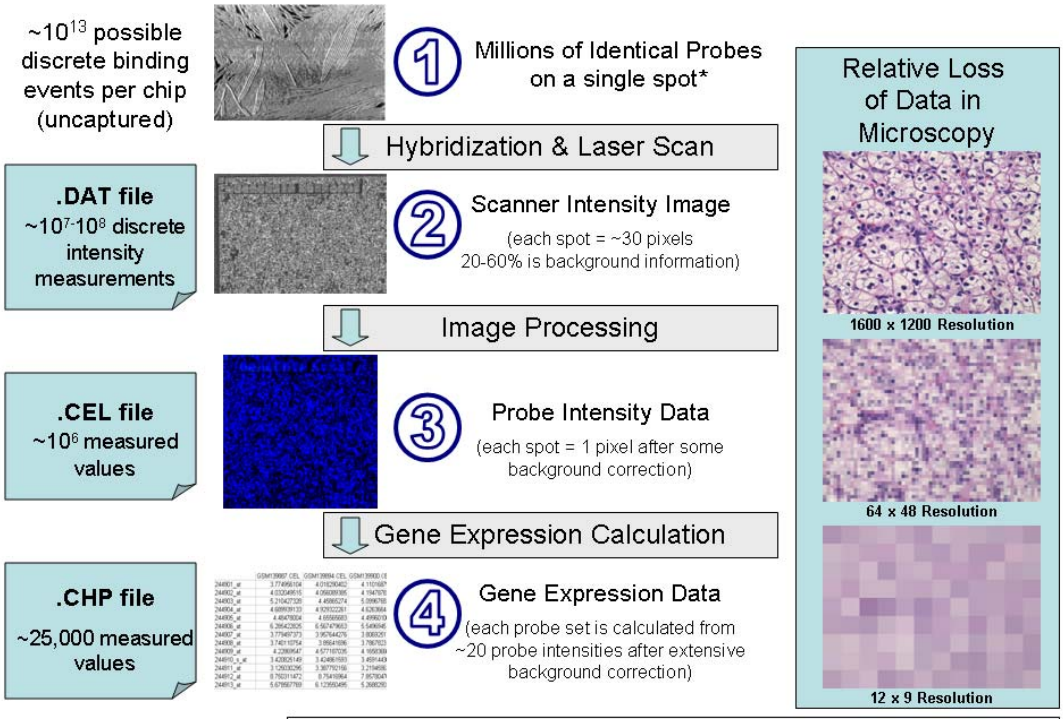


Figure 11: Diagrams showing the loss of data and precision during microarray processing. * Electron microscopy image of a microarray (adapted from reference [83]).

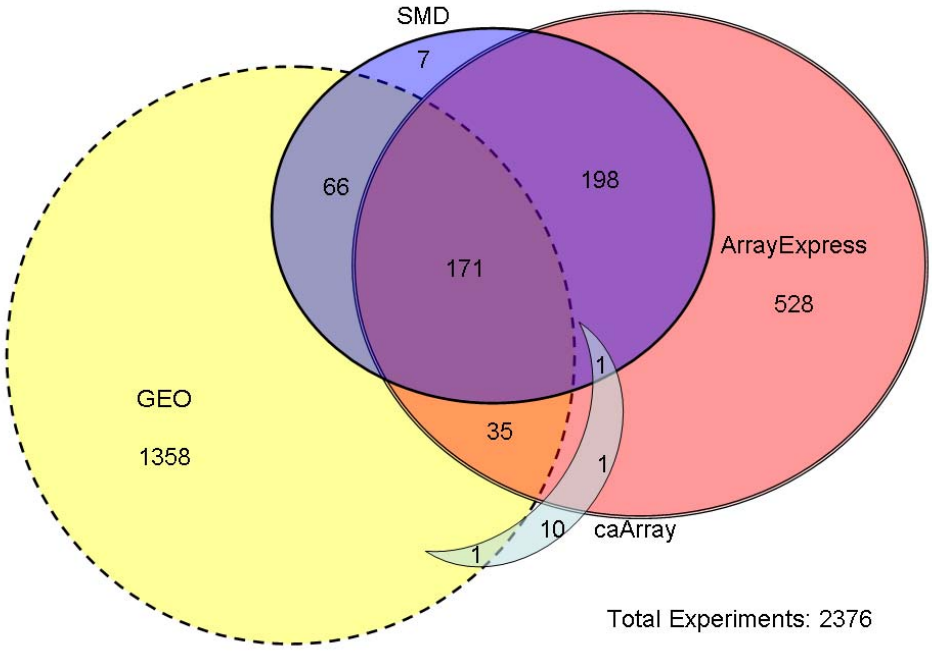


Figure 12: Venn diagram showing overlaps in experimental data between repositories.

Usability of Existing Repository Interfaces

Designing easy-to-use and clean interfaces to assist data providers and data consumers to upload and download is critical for expanding the reach of microarray repositories. However, the usability of existing repositories is harmed by a lack of shared standards for providing minimally required experimental data.

First, existing repositories have different requirements for data submission, and vary in their degree of openness to community involvement. These repositories vary in their ease of use of human interfaces (e.g., web sites), in the availability of programmatic access through Application Programming Interfaces (APIs) and their availability of data for bulk download (e.g., the entire database available through file transfer protocol (FTP) or export of search results in Extensible Markup Language (XML) format). In general, GEO and ArrayExpress have good web sites and APIs, but any effort to merge their datasets (as in ArrayWiki) requires developers to learn a variety of interfaces (Custom XML and SOFT files for GEO, MAGE-ML and seven custom file formats for ArrayExpress). Being the earliest developed repository, SMD does not make use of recent advances in usability such as JavaScript and AJAX. However, its functionality has been updated over time based on feedback from users and thus is far better than caArray, which is slow to respond and does not provide advanced search functionality.

Second, the existing repositories have different policies with regard to the timeline of making uploaded data available for public consumption. In some cases, this is a service to authors to allow them to use processing tools while keeping data private until publication. For example, GEO's express policy is to make data public automatically after six months. Existing repositories also have different data verification and curation because the database administrators vary. Some repositories will exchange emails with individuals making submissions to check facts. Regardless, at the detailed level of raw probe data, many problems still make it into the final repository, including corrupt file

formats and missing probe intensity files. Many experiment records claiming to include 200+ chips may only contain half that many files in the associated compressed data file.

Finally, the existing data repositories do not provide scanner intensity data, even though this data is extremely useful for quality control procedures. This data type absence certainly confounds down-stream data analysis because the artifacts caused by instrument and experimental procedures cannot be double-checked by the users.

Data Maintainability of Existing Repositories

Meta-data in existing repositories are usually problematic due to lack of standard in data maintainability design. One category of problem is the lack of meta-data. Most of the repository query interfaces are optimized for finding specific experiments from the literature, which is the first step taken by clinicians or biologists. (Based on the comparison survey we conducted, connections between experiments and PubMed are usually accurate.) However, they often do not provide technical features, i.e., meta-data such as number of samples, quality control measures, and probe-to-gene conversion methods (e.g. GCRMA or PLIER in Affymetrix technology). These features are critical for the downstream gene ranking and interpretation. Also, they often fail to provide and the correct dates of the experiments, and the associated protocol information. For example, some inaccuracies are a minor nuisance, like an experiment in SMD performed by Hong Juan on 11-16-1001 (instead of 11-16-2001), but others are more serious, like the problem in caARRAY where default experiment dates are all set to 01-01-2000, making that appear to be a wildly popular day to run a microarray experiment.

The issue of assigning an experiment date is unresolved in itself. Most microarray experimental results use arrays processed over a period of weeks, months, even years in some cases. When a data provider is expected to provide that field, they often just enter the publication date of the final paper. This is completely different from the timestamp of the data in the original arrays (e.g., the Affymetrix intensity data format contains a

timestamp that the array was scanned). Until now, no microarray repository has attempted to extract and provide that data.

Another category of problem is the lack of adaptability of meta-data. That is the adjustment of meta-data based on the evolution of Microarray data standards. Before widespread adoption of the MGED Object Model (MGED-OM), microarray repository designers were left to invent their own labels for each column in their database. This led to a lack of agreement in what is appropriate to make a required field, and what meta-data (data labels) make the most sense to users. caARRAY is the only repository based entirely on MAGE-OM standards, but it's impossible to map experiments to their meta-data using the current search interface. Based on all the issues discussed above, we design and develop a Wiki repository that can evolve meta-data standards at a rate the community demands.

Methodology and Development of ArrayWiki

An important consideration when creating a biological data repository is the reuse of data standards accepted by the community. However, there are only nascent efforts underway to standardize human curation interactions with data repositories [84]. Every repository still develops custom interfaces (usually web pages) for data access and modification. Technical experts might take the time to learn a specialized curation tool, but the wider community is unlikely to invest the time and effort. For this reason, the most difficult part of hosting a repository is recruiting and maintaining the interest among domain experts to contribute information and validation. The policy of many repositories of only allowing original data providers to modify their records adds to this problem (see Figure 13). The result is that while data is becoming increasingly sharable, it is also becoming increasingly stale [20].

The Wiki paradigm will likely be an important technology for data curation for biomedical research [85, 86]. Inspired by the spectacular success of the Wikipedia project

(http://www.wikipedia.org), there have been efforts to compile biological knowledge in a Wiki format [87-89]. Also, there have even been suggestions that the whole of medical knowledge may one day be accessible through this format [90]. These efforts are largely motivated by the ease of use of Wikis and the ready availability of high quality, free and open source wiki software, such as MediaWiki (http://www.mediawiki.org). Wikis provide readable information for both humans and computer programs (see Figure 14 and 15). In fact, recent publications have already shown that semantic web technologies such as automated annotation using Wikipedia pages have had some successes [91].

System interoperability efforts such as MAGE-OM [92], SBML [93], BioPAX [94, 95], and caBIG [96] rely on XML for machine readability. However, translation of XML into human-readable format is not a trivial process. The dbpedia effort (http://dbpedia.org/docs) is an open source project with the goal of automatically translating Wikipedia entries into the Resource Document Framework (RDF) format, which is a more recent and more flexible technology based on XML for Semantic Web. The Wiki syntax does not have a standard parsing structure like XML. However, the use of a smaller vocabulary of formatting syntax and "templating" improves the machine readability of its contents over that of typical unstructured web contents.

Wikis hold their greatest promise in the dramatic advances over XML in human readability. Future research may be able to fully integrate human-readable (Wiki) and machine-readable (XML) technologies [97]. Many users have the opportunity to modify Wiki data, and eventually consensus can be reached naturally. Many standards bodies (e.g. the SBML Consortium) already use Wiki software to accept community input before freezing a specification document. The BiomedGT Collaborative Ontology Development Wiki was designed by the National Cancer Institute to facilitate development of new ontologies with input from the wider community. They are currently working on many emerging domains, including nanotechnology, nutrition, biospecimens, and adverse event reporting.

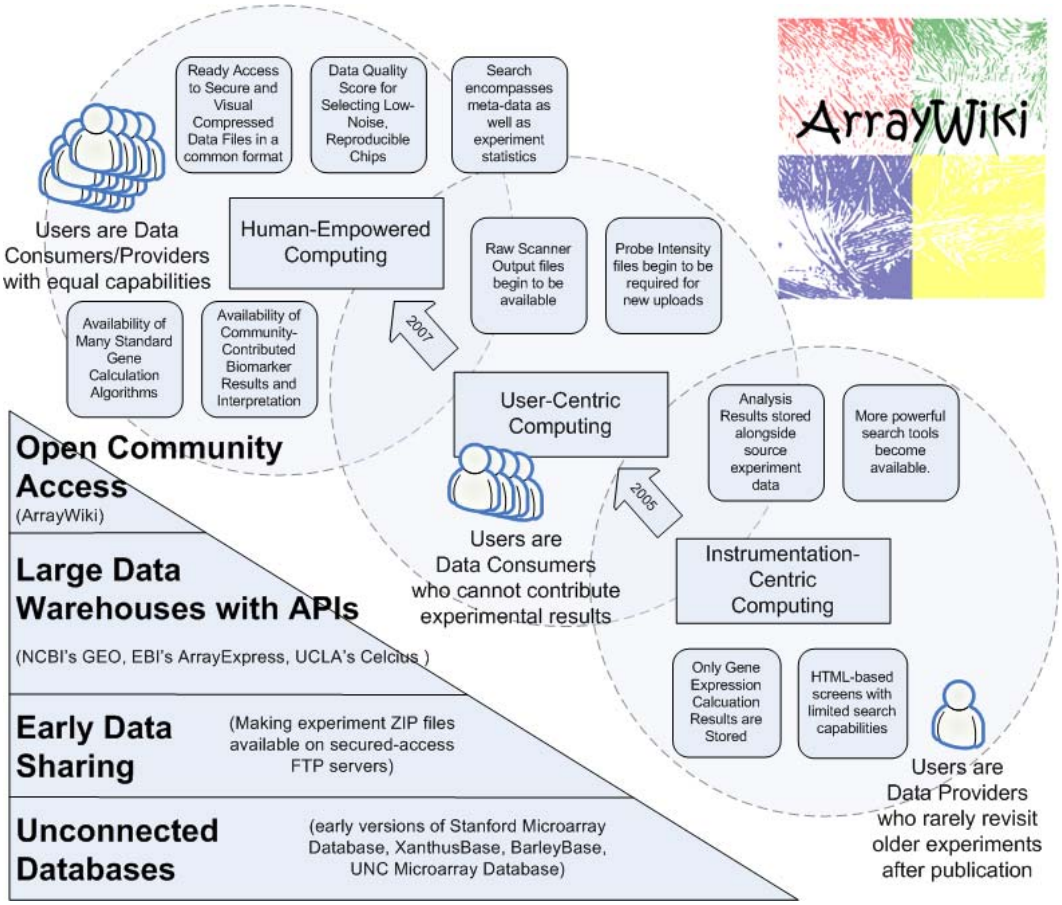


Figure 13: Evolution of biological data repositories (microarray case study). The capabilities of data consumers and data providers have changed over time. ArrayWiki represents the most open model, where all users can update data annotations and provenance is managed through the Wiki software.

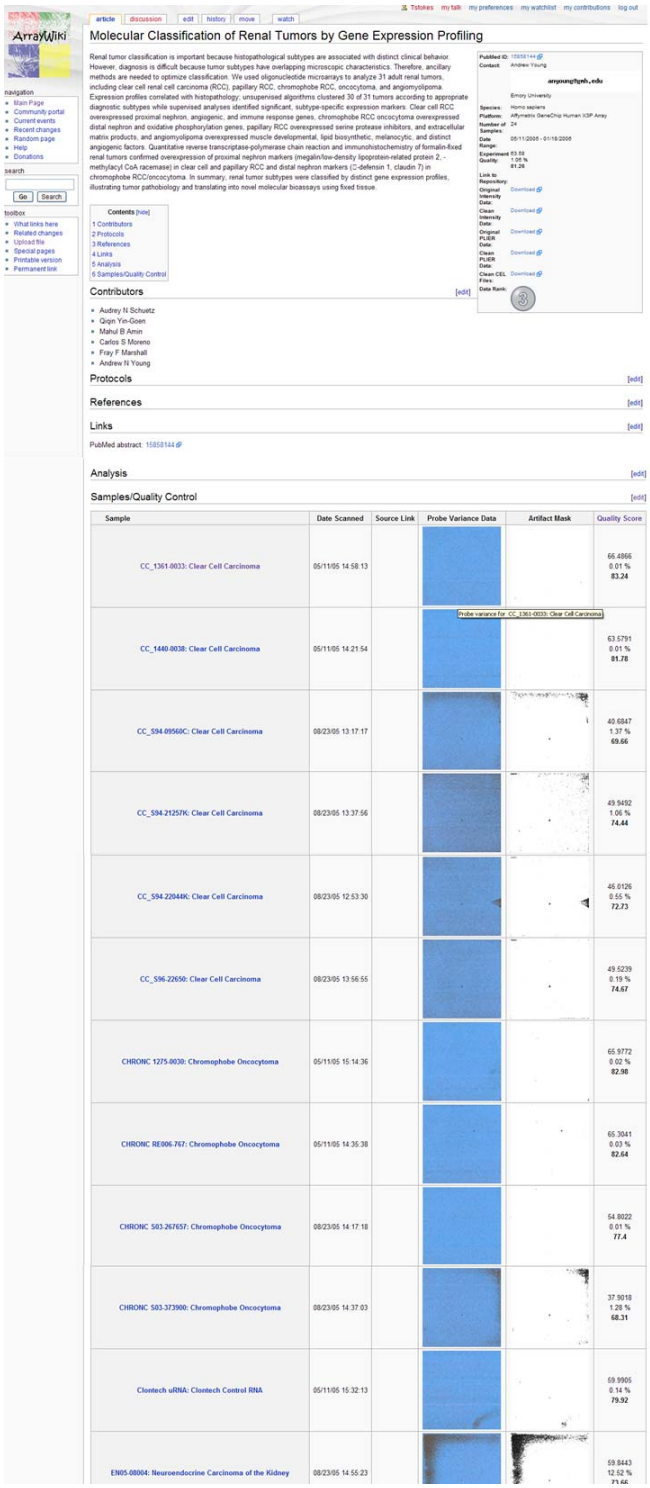
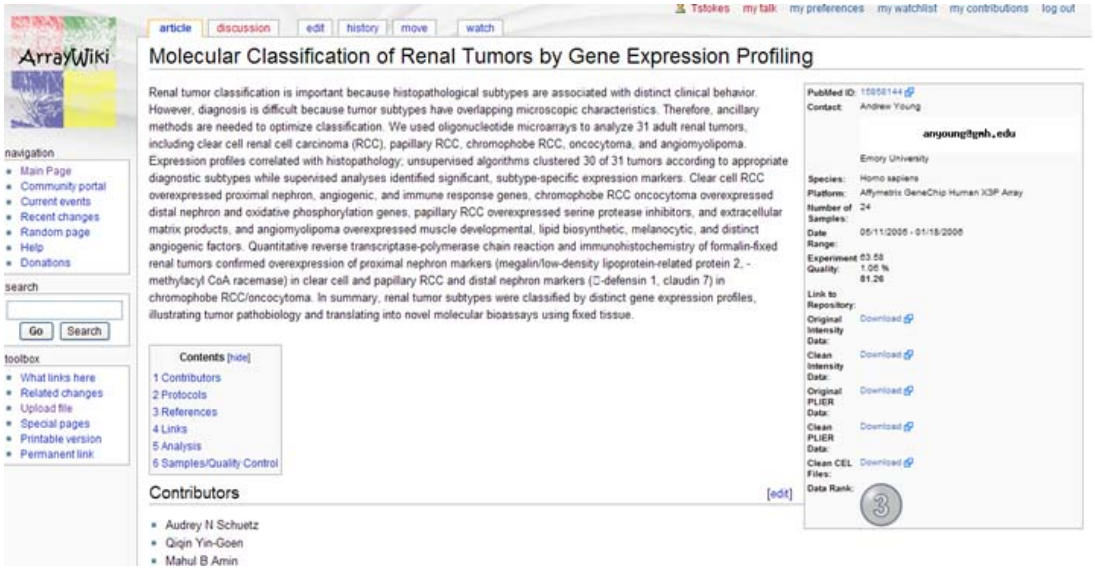


Figure 14: Sample experiment page in ArrayWiki. General experiment information such as contact info, data download links, and experiment summary are at the top. A table of samples at the bottom provides quality details and links to information about each individual sample, including BioPNG images of standard deviation data and scanner data.

A.



B.

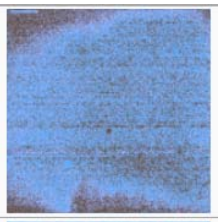
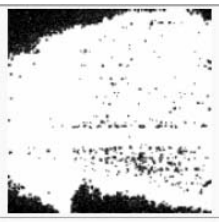
HL60 perturbed by 7.8e-06microM of chlortetracycline	11/14/06 12:17:10				76.7968 15.68 % **80.56**
HL60 perturbed by 1.48e-05microM of lidocaine	11/14/06 14:42:37				89.3828 0.01 % **94.69**

Figure 15: Detailed look at components of the experiment page. A) Representative experiment header in ArrayWiki. B) A close-up of two chips on the main experiment page with differing artifact patterns and quality scores, although they were processed in the same lab within 2.5 hours of one another. Meta-data uniquely available in this resource are the sample class assignments, quality scores, chip variance and artifact visualizations, scan dates and times, and visualizations of the NPIXEL and STDEV quality data generated by the array scanner.

System Design

In view of the limitations of repositories closed to community maintenance and the valuable features of Wiki knowledge repositories, we have developed ArrayWiki to host microarray data in an open environment, modifiable by any user. The culmination of ArrayWiki might be to unite data from other repositories, while providing the most detailed raw data and results of the latest best-of-class analysis algorithms.

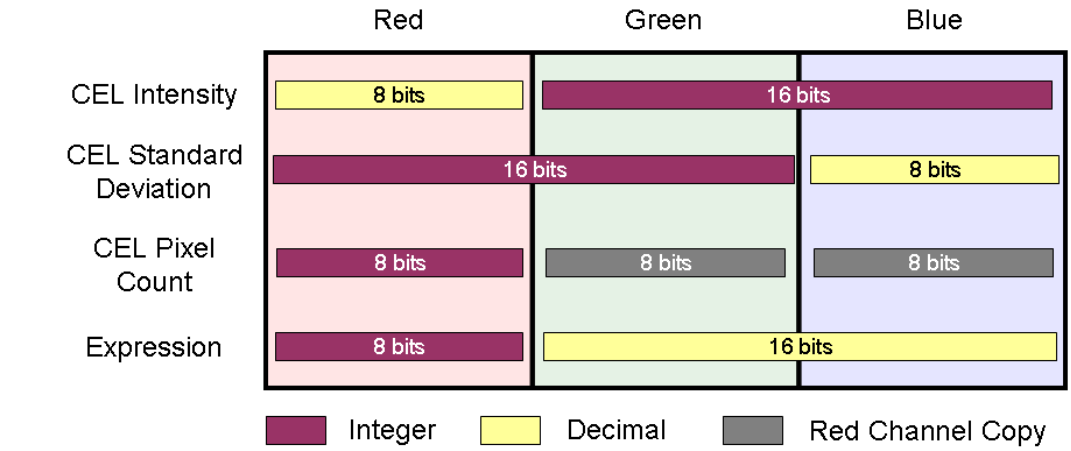
ArrayWiki pages are initialized programmatically by accessing APIs of GEO and ArrayExpress, or manually when an experiment does not exist in any repository yet. The current version contains over 650 experiments imported by GEO API (see Table 3). Quality control processes are still being run on these experiments to complete the import. A local database listing of all imported experiments ensures that existing pages are not overwritten each time the import process runs. A PHP class called Snoopy allows the import program to manipulate Wiki pages using HTTP POST, mimicking the process by which human users add contents. This is better than direct insertion into the database because it preserves the page history and the update tracking system, allowing for rollbacks of unintended changes.

ArrayWiki makes use of many useful add-ons to the MediaWiki software to enhance security and interoperability of data. One of these add-ons is the Captcha graphic for reducing automated spam generation. This feature requires the user to type a nonsensical word displayed in an image file whenever they add external links to a wiki page. Another add-on is the email image convertor. Contact emails are displayed as images in ArrayWiki to prevent mass harvesting of emails by automated scripts. We are considering a polite automated process that will ask data providers to update their contact information and flag email addresses that are responsive as an additional indication of experiment quality.

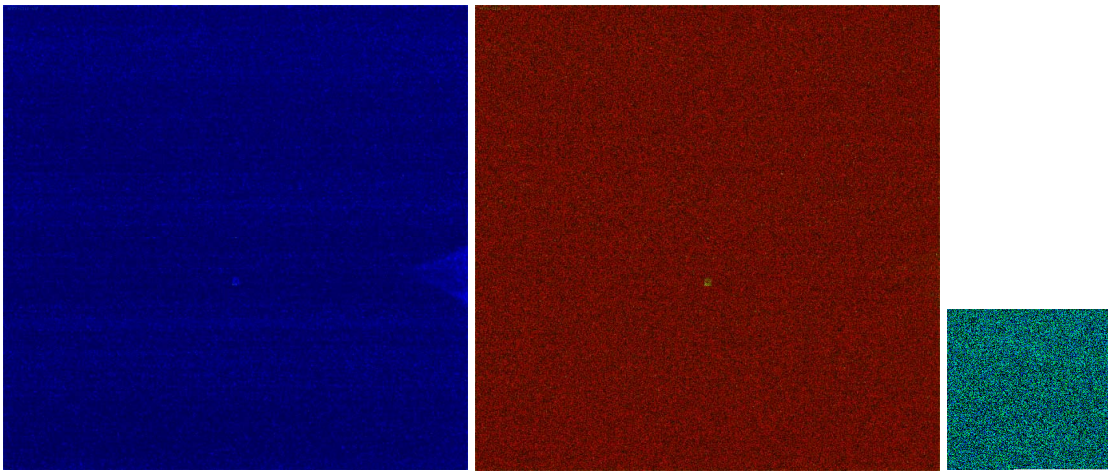
In addition, the import process accesses raw data files and converts them to the BioPNG format (see Figure 16,discussed further in chapter 4). This efficient storage method allows our system to support a greater data load and to make more efficient use of network bandwidth for downloads. This format offers greater protection against malicious software than ZIP files (which may contain embedded executables). Custom scripts have been added to convert BioPNG files into the Affymetrix version 4 CEL files to enhance data interoperability. These files are temporarily made available for download by clicking the link and later deleted to conserve file system space.

The import module of the ArrayWiki design is a multi-step process with many integration points and potential failure points (see Figure 17). We are continuing in our effort to make the process automated, especially detection of errors and error handling (e.g. splitting experiments that contain multiple chip platforms into separate Wiki pages). The import has been stress tested by importing 7029 microarrays from the Connectivity Map Batch release 2. Now, the seven steps of the import have been implemented as jobs (using the cron scheduler in Linux) that initiate every hour, allowing for as many as 24 experiments to be imported per day.

ArrayWiki runs a number of automated quality control processes during import and all results are stored on the page. In addition, the capture of novel meta-data allows for quality overview results as in Figure 18. It is recommended that users download "clean" data when available. The import program uses a standard Table or Infobox template for all meta-data to improve machine readability. ArrayWiki also uses templates (in the style of Wikipedia) that allow Semantic Agents like dbpedia to better interpret structured data. Over time, ArrayWiki may prove to be a useful tool for reaching community concensus on data specification and curation standards.



A.



B. **C.** **D.**

Figure 16: Examples of seucre BioPNG compression of Affymetrix CEL file data in PNG format. **A.** BioPNG distributes the components of a floating point numerical value into three color channels of a PNG image by first converting to binary representation and then splitting the string of bits into integer and decimal parts. The allocation of integer and decimal parts is determined by the precision required to store the data. The inverse relationship between CEL Intensity data and CEL Standard Deviation data causes the overall effect to be red or blue (**B.** and **C.**). This causes the data types, which might easily be confused when presented in text format to be immediately visually distinguishable. **B.** The Probe Intensity data file can also be used to verify problems such as edge effect. The original CEL files can be rebuilt using this file. **C.** The Standard Deviation data file gives a picture of scanner confidence in the measured values. Experimental protocol problems such as edge effect are usually seen here and support the variance heatmap. **D.** An example of the Expression data type, which has small integer values but very precise decimal values.

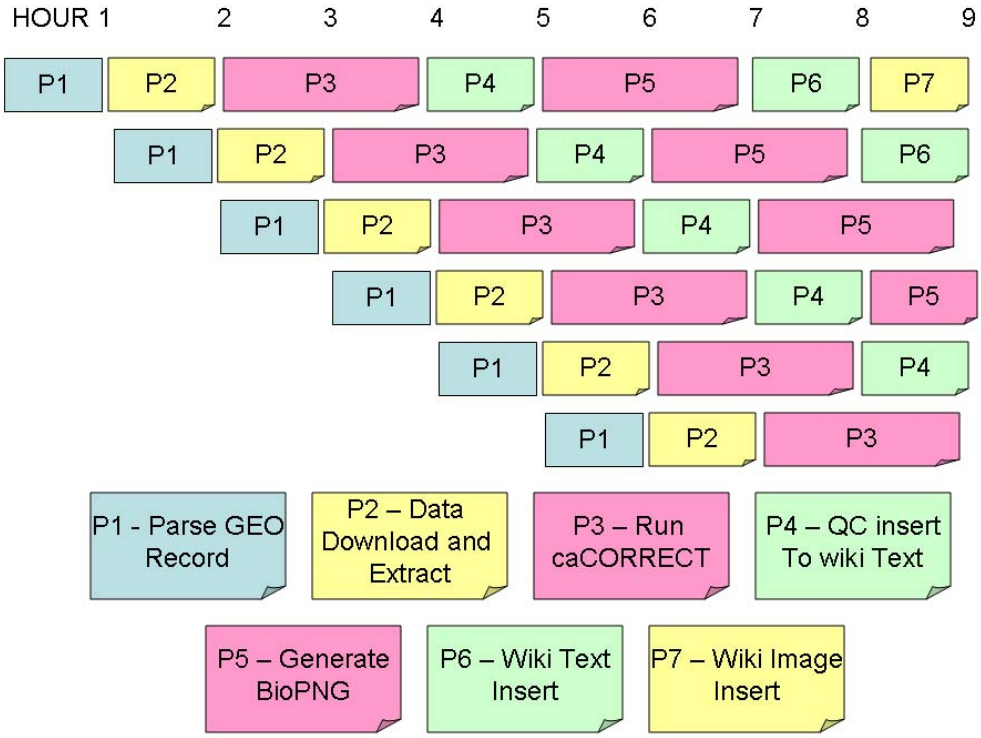
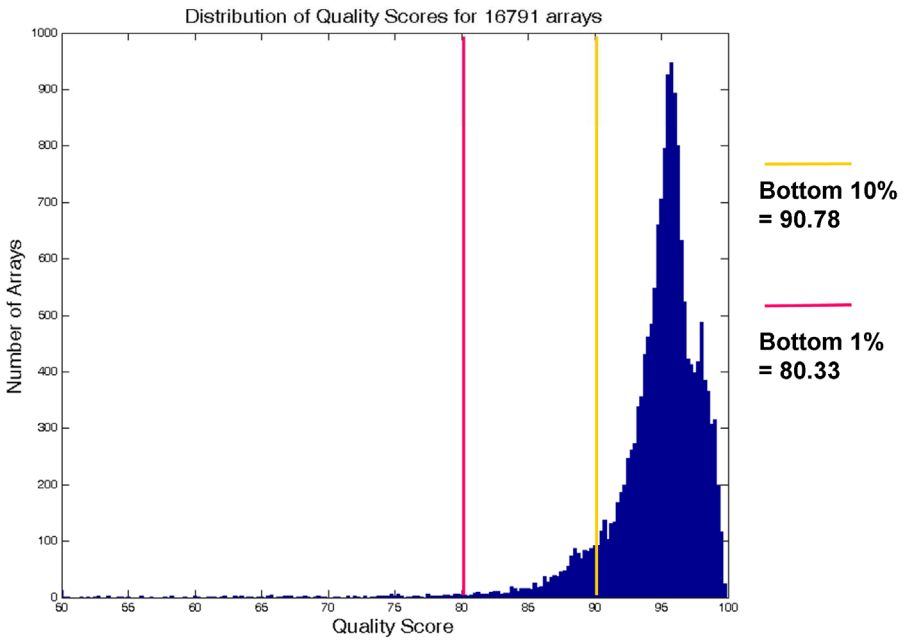


Figure 17: ArrayWiki automated import process details. At any given time, 7-9 jobs could be running. All of the check-pointing of data integrity between steps is managed using the ArrayWiki database. The import process was designed to be parallelizable and currently runs on two independent machines, both connecting remotely to the same database and inserting text into the production wiki site.

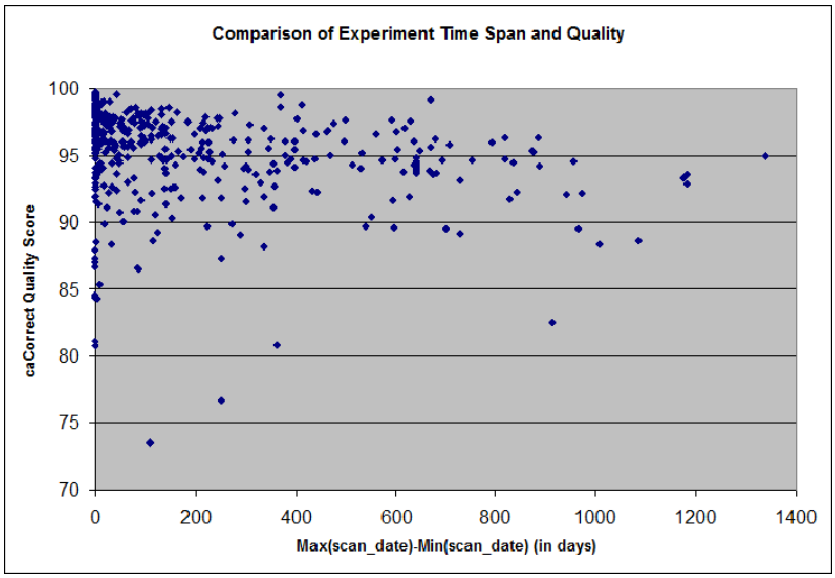
Table 3: Results of running ArrayWiki import over a period of six months.

Property	Value
Total Experiments Loaded	675
Experiments Attempted	1022*
Number of Chips Loaded	20,025
Data Sources (unique contact emails)	936
Average Number of Chips per Source	26.7#
Total Image Directory Size	68GB

* Causes for failure include: corrupt data files, missing chip definitions, and sample size < 4.
# Sample max of 200 may underestimate.
Connectivity Map (7029 chips from Broad Institute) excluded. (45.4)



A.



B.

Figure 18: Data quality features extracted using ArrayWiki meta-data. **A.** Histogram of 16791 microarray samples in ArrayWiki database. The red line at quality score 80 indicates the quality cut-off at which our team recommends discarding the sample from analysis. 1% of samples do not meet our quality requirements. Analysis using these arrays is seriously compromised. Most samples with quality scores between 80 and 95 contain artifacts that can be corrected by caCORRECT. These samples often are not identified as problematic by traditional QC means. **B.** Comparison of how experiment time span is related to quality. Long-running experiments are much more prone to "batch effect" because of change-over in technicians running the microarrays and changes to the laboratory environment in which microarrays are prepared and scanned.

CHAPTER 4

BIOPNG AND SCALABLE VECTOR GRAPHICS VISUALIZATION FOR BIOLOGICAL DATA

Biomedical Informatics has experienced an explosion in data analysis tools in response to the dramatic growth of data described in chapter 1. The biomedical researcher from a non-computational background is somewhat bewildered by this array of tools just as they are by the task of data management. The third objective of this dissertation is to develop an interactive web-based visualization system that enables fast and effective biomedical decision-making by integrating various data sources and by visualizing heterogeneous data and information. This aim addresses challenges of enabling efficient access to information in a fast-paced environment. This aim can be measured by how well the technology improves usability of tools. The key metrics of usability for TBMI are accessibility through the web, flexibility in visual rendering, and responsiveness even when handling high-throughput data. The key contribution of this research is that it only uses technologies that allow for embedding of the complete source data for the visualization into the same file that provides the visual data. This facilitates sharing and verification of data in complex workflows.

Biomedical data visualization is a great challenge due to the scale, complexity, and diversity of systems, interactions and experimental data. Standards for interoperable data are a good start to addressing these problems, but standardization of visualization technologies is an unsolved problem. SimpleVisGrid is a visualization system that builds on caBIG common infrastructure for cancer research, and clearly specifies and extends three standard data formats for inputs and outputs to the services: comma-separated values, Portable Network Graphics, and Scalable Vector Graphics. Four prototype visualizations are implemented: 2D array data quality visualization, correlations between high-dimensional data and meta-data, feature landscapes, and biochemical or semantic

network visualizations. The services and data model are prepared for submission for caBIG Silver-level compatibility review and can be integrated into automated research workflows. This work is in preparation for submission to the Engineering in Medicine and Biology Society conference.

Usability

Usability is the dominant design factor among the three discussed because it determines the usefulness of an integrated system and thus its impact on society (e.g. the scientific community and ultimately on patient health). Interoperability and adaptability increase usability, but also lend many practical advantages to the developers. Usability is the most easily overlooked by computational experts because it requires open communication with non-computational users to measure and improve. Usability also benefits from standardization, the most important catalyst for systems integration.

Most users (e.g. biologists or clinicians) have not been trained in computational algorithms [98]. Even for users that do have training, usability will continue to be critically important in bioinformatics system design. Bioinformatics systems often present unexpected results, but time is required to ensure that these results were not caused by a software bug. When processed results do not have an obvious derivation, the user has two choices: to blindly trust the results (which can propagate bad science when results are misunderstood and misused) or to look for another tool which will return something more sensible.

User acceptance is still a large obstacle for bioinformatics. One reason is that computational scientists do not place enough emphasis on the usability of their tools and underestimate the rather large time commitment required to install, configure, learn to use, and independently validate the results of their software. Possibly the greatest concern of biomedical researchers is the risk of system instability (PC crashes and resulting productivity loss) that accompanies any new software installation. In addition, many

computational tools overlap in functionality with established workflows because users typically develop their own non-computational solutions, only to search for a software tool after the manual process becomes unmanageable.

If bioinformatics algorithm experts want to capture advanced user feedback, they need to give users a sense that their needs are the priority by adding very simple usability feedback systems. Also, they should adopt web-based system development as the platform because almost every user is now familiar with how web browsers work and how to troubleshoot basic configuration problems. Lastly, integrated system developers can enhance user acceptance by forming interdisciplinary teams to increase usability and reliability of the system.

Usability Standards

Even very recent usability studies in the bioinformatics field are often based on high-level qualitative analysis. Some methods are under development, but none have reached wide adoption [99, 100]. Usability research requires investigation into the user's motivations for trying the system, how the user is rewarded for using the system, the culture of the user, the expectations that the user brings to the interaction including expectations for personalization of the experience, and understanding of the context in which the user is working.

One of the key problems in creating motivating visualizations for biological research is related to this understanding of context. Users with different specializations want to see lower levels of detail for concepts that they understand and higher levels of detail for unfamiliar concepts. This problem is often addressed in web applications by providing summary text reports with hyperlinks behind many identifying fields. Today, new technologies for providing scalable, interactive graphics will allow web applications not only to be even more portable to mobile devices, but also more visually appealing as well.

Systems Biology Graphical Notation (SBGN) Standard

Systems Biology Graphical Notation (SBGN) standard [101] is a collection of over 40 symbols for the presentation of biochemical reaction networks. This notation standard includes the ability to compartmentalize the model, to reuse logical blocks, and to represent ambiguous or poorly understood relationships in the model. The small circles attached to reaction species blocks represent single-site modifications such as phosphorylation.

The purpose of having a graphical notation standard is to add rigor and consistency to diagrams that convey biological network structure. Once the community has adopted a standard such as SBGN, its use will become familiar to everyone and the time spent on interpreting published diagrams will be reduced. A comparison to electronic circuit theory is appropriate here, because the graphical notation used for digital circuits has allowed large teams of engineers to collaborate in order to design and troubleshoot today's highly complex microprocessors. Most computational systems biology experts agree that this degree of complexity is the future for biochemical network models.

Gene Ontology (GO)

The Gene Ontology [102, 103] was originally designed to address the pace at which biological elements were being described relative to the pace at which they were being discovered through genome sequencing. Three ontologies were developed. The Ontology of Molecular Function describes what a gene product (a protein or a strand of RNA) does at the most fundamental biochemical level. The Ontology of Biological Process describes what a gene product does in the context of the biological objective. The Ontology of Cellular Component describes the location of the gene product within the cell. The gene product may not necessarily be a part of the makeup of the given component; it may simply be involved in an activity in the location described by a given

component. Most researchers refer to GO as the ontology of gene function, because the biological process and cellular component terms also support cellular function.

GO is represented as a directed graph of relationships and has a very limited set of relationships, currently limited to 'is a' and 'part of'. The visual representation of these relationships provides a very good picture of the context of each term in the ontology. However, these pictures can also be very misleading because there is no easy way to depict relative significance in the relationships between terms. Layout algorithms for the graphs are currently focused more on fitting the nodes into a given space than on giving spatial relationships some kind of interpretive value. GO might benefit by defining a degree of membership for each gene product to each term, which could help differentiate between primary, secondary, tertiary, etc. contributors to biological functions.

An analysis of the number of terms defined in the Gene Ontology shows a decreasing annual growth rate, from 27% in 2003 to ~10% as of January 2006, and contains ~25,000 terms as of March 2008. This indicates that the ontology is becoming more stable and that tools making use of the terms grow more reliable. However, the task of mapping gene products (proteins or active RNA) to all of the appropriate functional categories has only begun.

Performance Optimization

Performance (a.k.a. responsiveness) is a critical component to the usability of software systems. It is important that computational technology is used properly to save time for the human user, rather than becoming a hindrance to efficient completion of their work. However, it is a rule-of-thumb in software that optimization should happen last, after all other user functionality issues have been resolved. Benchmarking modules for timing, profiling algorithms for CPU and memory consumption, and properly applying best practices for performance requires extensive knowledge of the underlying technologies and hardware architecture. We have applied these techniques to many

technologies, including Matlab, Perl, Java Servlets, PHP, Python and Ruby. Many web technologies, especially AJAX, Javascript, and SVG, have not matured to the point of standardizing best practices for performance. However, in the development of the web visualization tools presented here, we have conducted performance analyses of these technologies and apply those best practices that have been identified.

Notable Previous Work in Biological Visualization

Visualization of biomedical data encompasses a diverse set of domain specialties, computational platforms, algorithms, and invariant representations (i.e. symbols or glyphs). Data visualization improves the efficiency of the scientific research by speeding the discovery of hidden patterns that may indicate key scientific findings or quality problems in the data acquisition step that should be resolved before starting analysis. Additionally, many visualizations are intended to enhance clinical decision-making [104], and may be a critical accompaniment to the successful translation of new molecular data acquisition techniques to the clinic and the advent of personalized medicine.

Attempts have been made to organize visualization efforts into classes of general problems: including network layout, clustering & correlation, and general images and plots (see Figure 22). The classification system presented here has already been reduced to visualizations of "non-tangible" data, and excludes a whole body of work focused on accurate representation of three-dimensional (3D) objects such as anatomical visualization and molecular conformation visualization. In general, the work of SimpleVisGrid has been to focus on the enormous field of two-dimensional (2D) data representation for interpreting data analysis results, with the minor exceptions of discussing the application of BioPNG to "data cubes" and the use of two-and-a-half-dimensional (2.5D) representation used to view feature landscapes.

Visual Statistical Data Analyzer (VISDA) [105] is the first visualization tool to become Cancer Biomedical Informatics Grid (caBIG) certified. Cytoscape [106-108] is a general network visualization tool that has deservingly received a lot of attention in this field. Cytoscape is an open-source standalone installation and makes some APIs available for other developers, but does not support web-based or grid-based requests. Haploview [109] is useful for comparisons of entire genomes to one another to look for small differences. GeneWindow [13] is an interactive Scalable Vector Graphics (SVG) interface that enables the user to browse gene sequences with a variety of annotation overlays. Matrix2PNG [110] is a simple and useful tool for quickly converting data stored in a matrix into a heatmap. There are some similarities between Matrix2PNG and the BioPNG system presented here. The primary difference is that BioPNG can be used for data transport. It always treats the encoding of data into a PNG as a potential two-way interaction, sacrificing some of the visual appeal of the graphic for the ability to retrieve the original data using only the BioPNG file.

Genome-Phenome Superhighway (GPS) OmicBrowse

The Genome-Phenome Superhighway (GPS) OmicBrowse [14] is a web-based genome visualization tool developed by the RIKEN Genomic Sciences Center in Japan. The tool is web-based, so there is no need to install or upgrade local software. GPS is integrated with 12 databases and provides text search on all of them simultaneously [15]. This tool demonstrates a focus on usability as every result is translated into a location on the genome of the selected species. Four species are currently supported.

The primary drawback to OmicBrowse as an integrated tool with wide appeal is that the interface was developed on a proprietary platform (Macromedia Flash, now owned by Adobe Systems, Inc.). For this reason, the connections to the back-end data servers are obscured from user inspection, so data currency is difficult to verify (recent visits to the GPS website indicate that this problem may have been overcome as this

group has made their software and database backend freely available for download). Additionally, the visual components cannot be incorporated into a larger suite of tools except as a complete package, forcing other bioinformatics groups to duplicate effort in order to implement genome browsing on a more open platform. Open standards now exist to improve the portability of visualizations.

Database for Annotation, Visualization and Integrated Discovery (DAVID)

DAVID [12] provides a set of data-mining tools that systematically combine functionally descriptive data with intuitive graphical displays such as interactive biochemical pathway maps, protein functional domain charts and Gene Ontology charts. DAVID is hosted at the National Center for Biotechnology Information (NCBI) at the National Library of Medicine of the National Institutes of Health. A user session in DAVID starts with uploading a list of differentially expressed genes from a microarray experiment. DAVID then connects to a number of public annotation databases to create an annotation summary, including:

1. GenBank - Accession number corresponding to the nucleotide sequence
2. Unigene - Cluster containing sequences that represent a unique gene
3. LocusLink - Unique and stable identifier for curated genetic loci
4. RefSeq - Reference sequence standards for mRNAs
5. Gene symbol - Official gene symbol included in the Locus Report provided by NCBI
6. Gene name - Official gene name included in the Locus Report provided by NCBI
7. OMIM - Catalog of human genes and genetic disorders
8. Affymetrix - description Probe set description provided by Affymetrix
9. Summary - Functional summaries included in the Locus Report provided by NCBI

10. Gene ontology - Controlled vocabulary applied to the functions of genes and proteins. Functional classifications used here are those included in the Locus Report provided by NCBI

Visualizations in DAVID are a combination of HTML pages and images generated on the server for each requrest. The biochemical pathway maps are pre-generated images hosted at Kyoto Encyclopedia of Genes and Genomes (KEGG) [111, 112]. There have been a number of advances in mapping these pathways centered around the KEGG database, including efforts to transform these maps into interactive SVG documents [113-115]. DAVID also incorporates many features of the GOMiner Gene Ontology mining tool [116].

MAPPFinder

MAPPFinder [117] is another integrated visualization tool that will render images of differential gene expression in the context of gene networks, the Gene Ontology, or biochemical pathways. MAPPFinder uses the Gene Map Annotator and Pathway Profiler (GenMAPP) workbench to allow users to define their own diagrams or pathways for presentation.[118, 119]. In addition to displaying the gene expression in a variety of contexts, MAPPFinder will identify patterns of gene expression correlation in the Gene Ontology. The diagrams and analysis results of MAPPFinder sessions can be shared online using the MAPPFinder web site archive. The recent release of GenMAPP 2 exposes many new visualization capabilities to the user. Pathways can now be translated between species using homology information. A new mode of data visualization supports analysis of complex data, including time-course, single nucleotide polymorphism (SNP), and splicing. GenMAPP version 2 also offers innovative ways to display and share data by incorporating HTML export of analyses for entire sets of pathways as organized web pages.

Problems of Data Scale

One reason that so many visualization systems are designed as standalone tools or designed to connect only to pre-defined local databases is the scale of the data involved. The idea of plugging visualizations of such large-scale data as 30,000 genes on a microarray, 100,000 proteins in an interaction network, or billions of base pairs on a genome into a grid-based workflow for multiple-sample comparison has been technically infeasible up to this point. One obstacle has been that interoperability standards such as MAGE-ML [92, 120] or BioPAX [94], while extremely useful for passing meta-data about experiments around, cannot be extended to raw experimental results because the uncompressed text and required tags are too bulky. A balance must be found between appropriate meta-data to transport in structured formats versus bulk data to transport in standard compressed formats. A recent move in this direction is MAGE-TAB [121], which emphasized the facility of moving data between human-readable spreadsheets and machine-readable data files as a major usability factor for microarray analysis tools.

Achieving Usability with Scalable Vector Graphics (SVG)

A little over a decade ago, there was a lot of excitement around applets when Java technology was first released. This architecture built around the small, highly interactive applications available through your web browser spawned many spin-offs, including Macromedia Flash Applications and Microsoft ActiveX Controls. Of these three, Java applets were the only applications that could be developed without buying proprietary development software. Unfortunately, they suffered from stability problems, highly restrictive security limitations, and could not be easily scaled beyond single-purpose modules. For most of the past few years, users were forced to content themselves with the simple “click – hourglass – complete page refresh” interactivity of HTML forms when using web-based systems.

In 2003 the Scalable Vector Graphics (SVG) specification was approved by the W3C and received native browser support. SVG documents describe 2D graphics in an efficient way and can preserve the underlying scientific data. SVG documents can be annotated with custom tags to provide extensibility and tighter application integration. All SVG documents are zoomable (multi-scale) interfaces. This means that the resolution of the display device (or the available screen space in an integrated application) does not affect the readability of the scaled representation of the graphic. Finally, SVG documents can be programmatically manipulated using the Javascript Document Object Model (DOM), providing a means to implement innovative interactive interfaces. SVG has an important drawback in the lack of a standard set of widgets (buttons, sliders, select boxes) that make user interface creation much easier. SVG-based bioinformatics tools are not common. Some examples are GeneWindow [13], ArrayXPath [122], the microbial genome viewer [123], caCORRECT [18] and GOMiner [116]. SVG could be the basis for multimodal scientific and educational material in the future [124].

SVG rendering is provided in Microsoft Internet Explorer 7 by means of an unsupported Adobe SVG Plugin but is natively supported in the latest releases of the Mozilla Firefox, Apple Safari, Google Chrome and Opera browsers. SVG is compatible with Asynchronous JavaScript and XML (AJAX) technology. AJAX is a technique for manipulating an HTML-based user interface using background browser processes without causing the noticeable page refresh of classic HTML forms. AJAX has the advantage of being similar in look and feel to locally installed software. However, AJAX without SVG must rely on reloading of server-generated images to improve the interactivity of visualizations. Web-based visualization technologies have many trade-offs between openness, performance, and capabilities (see Table 4). The choice of SVG as a standard for SimpleVisGrid was made because of unparalleled openness, flexibility for data storage, scripting, and availability on mobile platforms like the iPhone.

Table 4: Comparison of visualization technologies for the web.

Technology	Primary Purpose	Release Date	Development Cost Status
Adobe (Macromedia) Flash	Interactive or animated web content with wide range of complexity	1996	Adobe Director (~$300)
Java Applets	GUI container for Java applications in a web browser	1995	Free APIs available, no supported IDE
W3C's Scalable Vector Graphics	Open-source and plain text description of vector graphics	2003	Various Free IDEs (e.g. Inkscape) and supported by Adobe Illustrator
Yahoo! Pipes	Graphical assistance with building web workflows and mashups	2007	Free on the Yahoo! Site with registration, currently in "beta" status
Microsoft Silverlight	Competitor to Adobe Flash, which was perceived as the dominant technology	2007	Free
HTML5 Canvas Element + AJAX	Introduced by Apple Safari and has been slowly adopted by other browsers	2008	Free, No IDEs identified

BioPNG: Data Compression & Visualization

Portable Network Graphics (PNG) is an open specification for image compression. Since images are simply representations of 2D data matrices, it is somewhat intuitive to think of image formats as candidates for storing data in this form. However, we found no examples of systems making use of image compression to store non-image data. There are two reasons for this: (1) general compression formats such as ZIP have been the obvious choice for compressing any data and (2) image formats tend to have limited bit depth and/or lossy compression algorithms that meant the data you got out might not match the data you put in.

We introduced BioPNG, a method for converting floating point numbers into color channels for storage in a lossless format, as a companion technology to ArrayWiki [68]. This format has many advantages over ZIP: (1) it is natively supported over HTTP, (2) it is presented easily in browsers so the data can be "seen" as it's received, (3) the compression rate compares favorably to ZIP, often performing 20-30% better, (4) PNG files do not harbor viruses like ZIPs can, and (5) meta-data such as scale (x and y dimensions) are easily extracted without parsing. Figure 19 illustrates the trade-offs of different BioPNG formats. PHP Source code for BioPNG can be found on ArrayWiki: http://arraywiki.bme.gatech.edu/index.php/BioPNG_Source_Code.

The BioPNG algorithm was developed to allow ArrayWiki to scale up faster while requiring fewer storage and network resources. Compression of laser scanner microarray data has been addressed by Luo and Lonardi [125]. The authors stress the importance of lossless compression and compare compression results of JPEG-LS, JPEG 2000, PNG, and TIFF image formats. They recommend JPEG 2000 but concede that this format lacks common browser support on the web. They also suggest (but don't implement) separating header info, foreground, and background pixels. As a trade-off between good compression, portability, and ready viewing of data, we have found PNG compression to be the most convenient.

Compression of Array Data

BioPNG works by first splitting the numerical formats into coarse-grained and fine-grained bins (see Figure 19), and then making use of higher-order filters available in the PNG library to model the data and store only the errors in the model. Affymetrix microarray data contains many non-Gaussian correlations in the data that can be exploited for the purposes of compression. Our research has shown that different microarray platforms can differ significantly in the entropy (in the information theoretical sense of the term) of the data. We have calculated the first-order entropy of the

HG_U95Av2 platform, containing 409600 intensity values, to be 10.1613 bits based on the samples we've processed. This means that an optimal first order compression algorithm should create files of average size 520.25 kilobytes (kB). By comparison, BioPNG compresses this data into files of size 767.42 kB. Including more chips in the calculation of entropy will certainly raise this estimate, as not all of the available intensity symbols were used in our study. Our results indicate that BioPNG compression performs better than any custom first order compression algorithm, while still providing good portability and visualization.

Most repositories provide gene expression and probe intensity data in a zipped format. This can be problematic when attached to emails and may be infected with malicious software by anonymous sources when shared on the web. BioPNG encodes Affymetrix probe data at 2.26 times compression over GEO's method of zipping each binary Affy file individually and then zipping all of the files again into one file. BioPNG's level of compression comes at small performance expense and no loss of data from the most important probe intensity measurement. Data stored in the file header are automatically transferred into the experiment metadata in ArrayWiki.

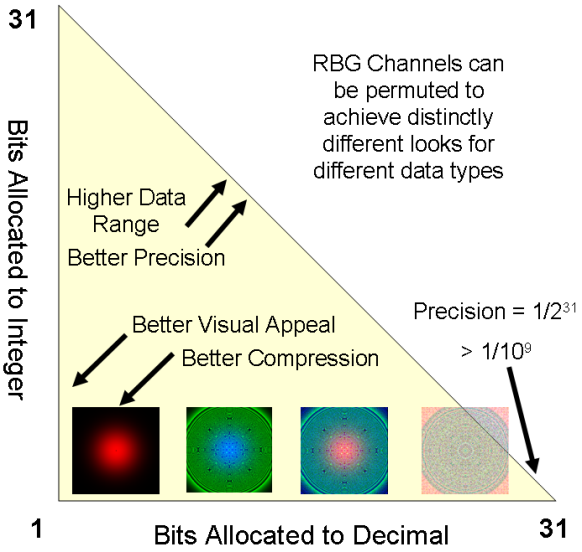
Individual probe data is impossible to measure with current scanning technologies, but this is likely to change with technological advances. As the space savings indicated in Table 5 become greater, the coding effort and computational expense to extract specific data points out of the files become greater. Retrieving a small set of intensity values from BioPNG is much faster because only those selected pixel values need to be converted.

Table 5: Microarray data storage formats and relative compression ratios.

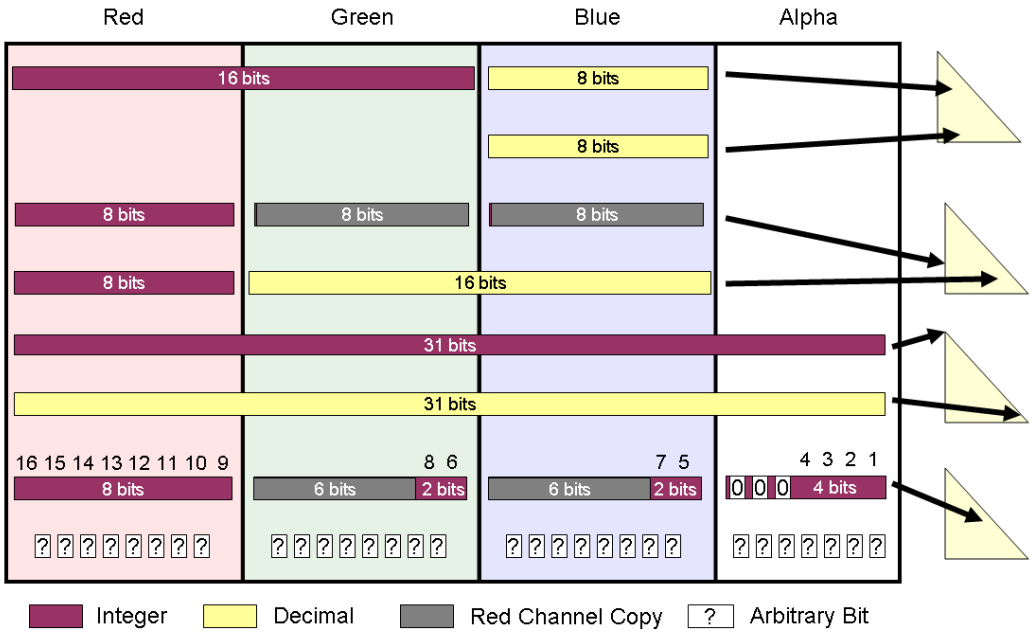
Scanner Intensity Image (Avg of 4 files)				
Native Type	Dimensions (pixels)	Size (KB)	Size as PNG (KB)	Space Savings
TIFF	1926x1974	7428	729	90.2%
GIF	1877x5486	9358	7975	14.8%
DAT	2920x2920	16654	10786	35.2%
DAT	8661x8661	146533	89497	38.9%

Probe Intensity Data (Avg of 4 712x712 CEL Files)				
Data Format	FileSize (KB)	Binary	Double gzip'ed	BioPNG
ASCII Text	11,683	57.6%	81.8%	91.9%
Binary (Affy v4)	4953	-	57.0%	81.0%
Double gzip'ed	2130		-	55.8%
BioPNG	942			-

Gene Expression (Avg of 4 CHP files, 22812 probe sets)				
Data Format	FileSize (KB)	Excel native	Zipped CSV	BioPNG
Excel native	2718	-	70.6%	91.6%
Zipped CSV	799		-	71.5%
BioPNG	228			-



A.



B.

Figure 19: Illustration of BioPNG encoding. BioPNG specifies encoding methods for many types of numerical formats into the color channels of a PNG file. **A.** The trade-offs to be considered between precision, data range, visual appeal, and compression when selecting an encoding scheme. Encoding a value simply requires converting to a binary decimal and splitting the resulting string into bit lengths to fit into the color depth required to store the data. The images along the bottom are all Gaussian surfaces stored at various decimal precision. PHP Source code for BioPNG compression and extraction can be found at the url: http://arraywiki.bme.gatech.edu/index.php/BioPNG_Source_Code **B.** The seventh line shows the encoding scheme used to encode 16 bits for Gel Plots, producing a grayscale effect. The final line of question marks the custom line is to show that the BioPNG API supports arbitrary definitions of color encoding.

Visualization of data distributions, data errors, and quality problems

The BioPNG data format provides features for ensuring data quality in addition to providing compression and protection from malicious software. The ArrayWiki import process generates a histogram of the original and the clean intensity data. This histogram stores the counts for all 490,000 possible values for intensity measurements in unprocessed CEL files, and the corresponding counts after the artifact removal process. Viewing this file can indicate data problems if single values are strongly over-represented or if an unexpected periodicity is observed in the data. Another histogram image stores the probability density function for each of these values, which is simply the hit counts normalized so they sum to 1. These images are useful in calculating the first- and second-order entropy of different microarray platforms.

The Gel Plot visualization is another application of the BioPNG format to the data quality problem (see the example in Figure 20). The Gel Plot is created by first converting the intensity values into the log10 space, and then binning the values into 600 bins. Like other BioPNG formats, the decompression method can perfectly reconstruct the log10 intensity counts used to create the file.

Quality problems may arise in the algorithm that converts the intensity values read by the scanner (available in the DAT file) into the values reported in the CEL file. The final effect on reported gene expression has not been quantified, but the extent of the potential problems is visualized by the BioPNG-formatted NPIXEL file. This column of data in a CEL file indicates how many pixels were used in the calculation of the intensity. This number generally ranges from 12 to 36. Systematic patterns in this image may indicate misalignment between the track of the laser and the grid of the microarray spots.

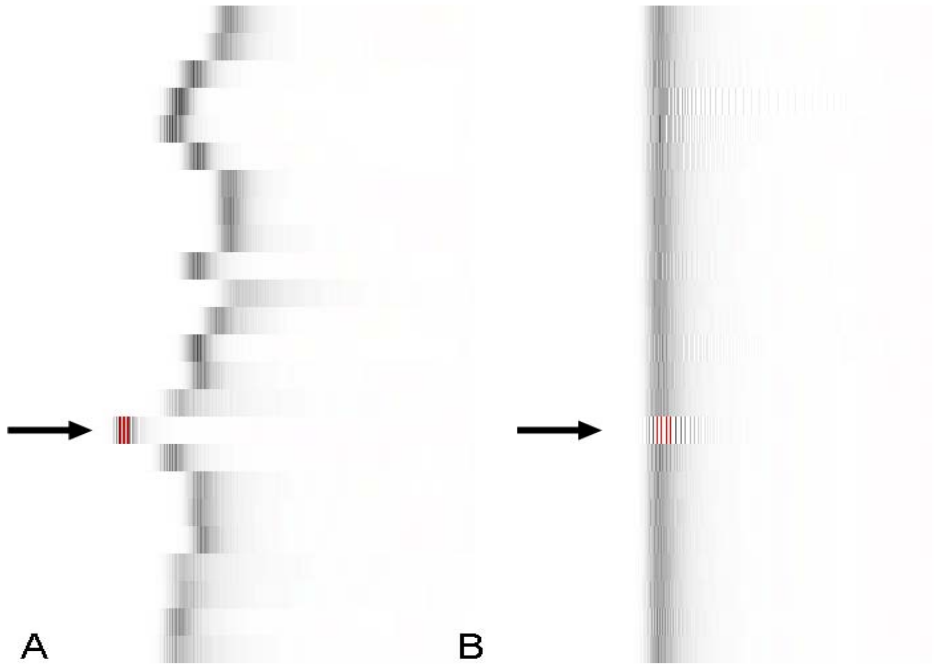
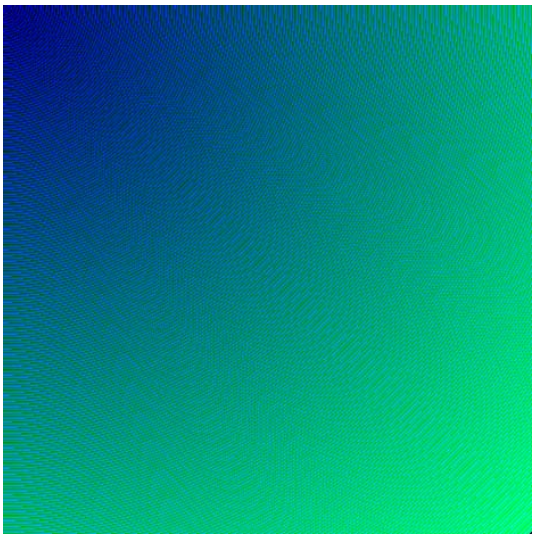


Figure 20**:** The BioPNG Gel Plot. (A) The gel plot of the original scanned data for the Renal Cell Carcinoma Affymetrix X3P GeneChip. The arrow indicates a chip with a corrupted file format, causing the intensity values to be read incorrectly. The Bioconductor program that parsed the chips did not catch this error, but it became clear after visualizing the intensity distributions. (B) The gel plot of the normalized and artifact-replaced data output by caCORRECT. These distributions are clearly well-aligned, but the same data parsing problem persisted without causing any post-processing algorithms to generate errors.

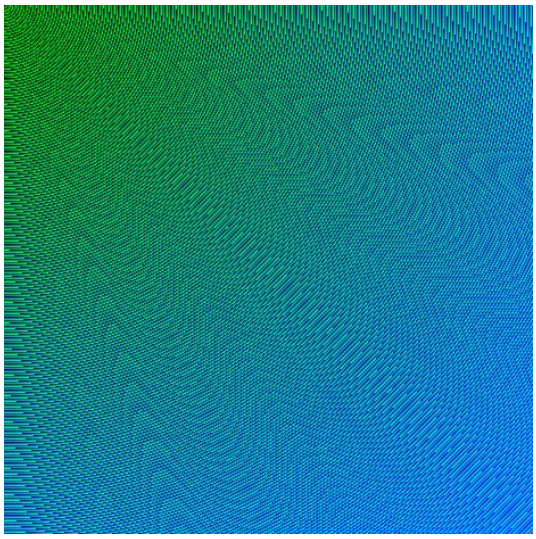
Extension of BioPNG to all 2D Data Types

Partly inspired by Matrix2PNG [110], we extended BioPNG to support multiple data types and to allow the user to arbitrarily choose color channel assignments in order to help users differentiate between different data types (see Figure 21). An interactive web site was developed to allow users to experiment with different options, and to show in what cases data can be completely recovered and in what cases some data will be lost.

A.



B.



C.

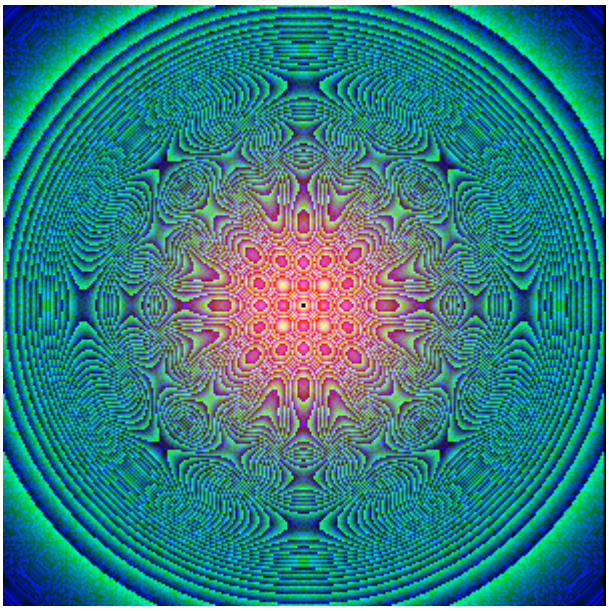


Figure 21**:** Examples of BioPNG Compression output using synthetic data. A. and B. This data file is a linear gradient from 0 to 1. The data precision makes only two color channels necessary, so red is not used. In (A), the blue channel encodes the lower order bits and in (B), the green channel encodes the lower order bits. C. This data file is a Gaussian curve starting at 1 in the center and dropping to 0 at the corners. While many visual artifacts do arise due to discontinuous bit incrementing, the overall smooth circular trend can be observed. The precision of this data requires all three channels to be used for complete reconstruction of the data. Data displaying gradual trends such as these compress very well using the PNG algorithm.

SimpleVisGrid: Visualization Services for the caBIG Community

With a wide variety of available visualizations, a necessary first step in the design of a real working system is to narrow the range to those especially suited for biological knowledge. Figure 22 shows one possible classification system. The main categories are Networks, Correlations, and Images. A large class of simple plot visualizations like scatter plots and histograms could be considered a fourth class, but I group those under correlations to simplify the system. Images were discussed in the previous section, with the exception of a technique using the Animated PNG (APNG) specification to encode data cubes into a series of PNG files that will animate using a browser like Mozilla Firefox. Though this technique is available in SimpleVisGrid and offers many similar advantages to BioPNG, the compression is less than ideal because each 2D file is compressed in the standard way and no additional filters are run in the third dimension.

This section will highlight network and correlation visualizations. The network visualizations work as a wrapper around the GraphViz library of algorithms, and so support a great deal of flexibility with improved visual appeal using SVG. Only the most novel correlation visualizations are presented here: feature landscapes and meta-data correlations. Correlations may best be presented using either SVG or PNG, depending on the scale of the data that must be rendered. Most of the visualizations presented here use microarray data as a case study, but these services could just as easily be used to represent other high-dimensional data types, such as mass spectrometry experiments or Surface enhanced Raman scattering (SERS) experiments.

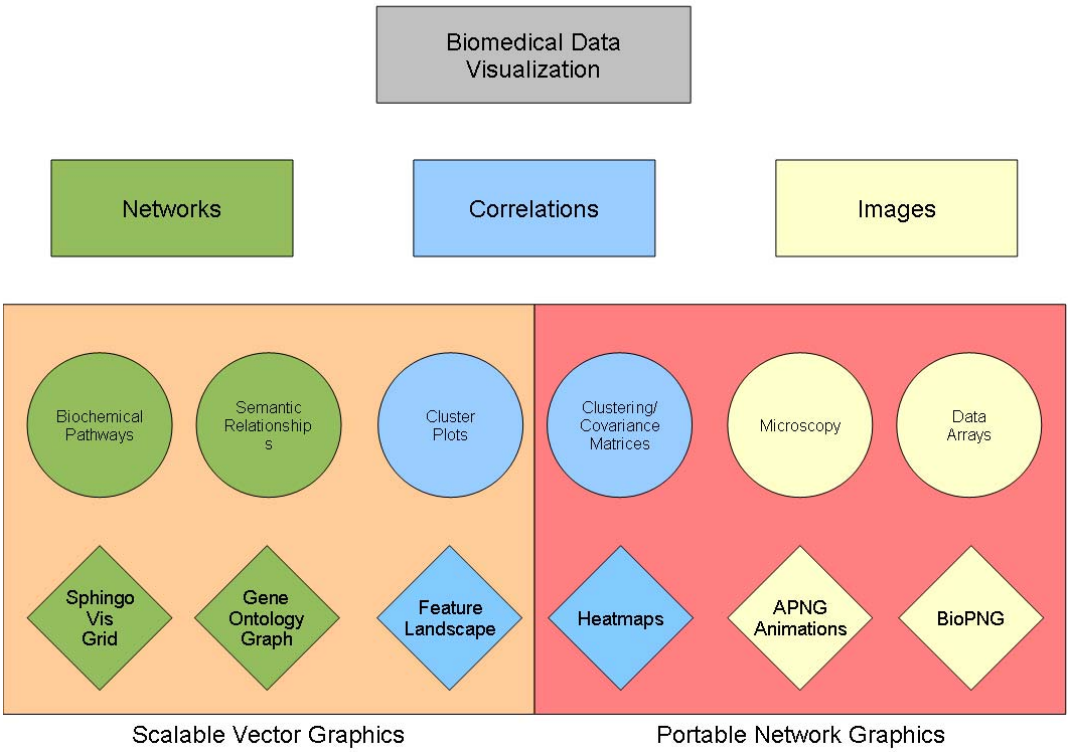


Figure 22: Classification of visualization technologies provided by SimpleVisGrid. The rectangles indicate the class, circles represent visual representations, and diamonds represent concrete examples of the representations. SimpleVisGrid supports only a subset of all of the possible methods to visualize heterogeneous data. The available visualizations are divided into three categories: Networks, Correlations, and Images. Six visualization types will be exposed, built upon two primary technologies: Scalable Vector Graphics (SVG) and Portable Network Graphics (PNG). All of the visualizations are 2D or 2.5D. In the case of the Feature Landscape, a single perspective is available on a 3D landscape. In the case of Animated Portable Network Graphics (APNG), 3D microscopy images are stored as many 2D images layered into a lossless compressed video animation file. Native support for APNG was added to Mozilla Firefox in version 3.0. Microsoft Internet Explorer does not natively support SVG or APNG, which is an obstacle to full adoption of data-embedded visualizations.

Visualizing Biological Network Relationships with Graphs: GridGOMiner and SphingoVisGrid

Graphs are a natural way to encode biological relationships. Three examples include the visualization and analysis of: (1) ontological relationships between gene function terms, (2) biochemical networks such as lipid synthesis, and (3) shared components of complex processes such as the Human Disease Network [126-128]. Additional examples include (4) genealogy and genetic pedigree of individuals [129] and even (5) the Tree of Life online phylogentic collection collaboration [130, 131].

Network layout problems frequently arise in biomedical visualization. While the problems can be very large, requiring a great deal of computational time to render, many are small enough to be rendered using simple algorithms. A good package for providing flexible and easy-to-learn algorithms is GraphViz [132]. Our network layout visualization service (see Figure 23 for an example) is a wrapper for GraphViz and provides higher-quality SVG documents than those obtained directly from GraphViz, including buttons to navigate among time series data and support for encoding information in the edges. We support many of the standard features of the GraphViz package in addition to automatic generation of a PNG file and embedding the original source data into the SVG document that is returned. In addition to biochemical reactions, the network layout service supports the layout of directed acyclic graphs (DAGs) like the tree structure used to represent relationships between terms in the Gene Ontology.

In developing a general framework for these visualizations, I studied eGOMiner for integrating multi-experiment GO functional analysis. Additionally, I adapted this analysis specifically for looking at drug response (time series) experiments in the context of a Human Disease Network [128] in a system called nanoDRIVE. I also deployed a visualization system with accompanying database, PathwayVis, which integrates data from PubChem, LipidMAPS, and quantitative mass spectroscopy experiments to display

molecular concentration measurements on complex biochemical pathways. This system supports comparisons between empirical data and simulation results.

These efforts have not been published, but improved my understanding of the diversity of data involved in building flexible grid services. The three tools have very different inputs, outputs, and interpretation goals, but they have very similar system components.

eGOMiner: Comparing Gene Functional Significance Studies

The GOMiner [116] tool has been used widely to interpret the results of microarray experiments [133-135] (the paper shows 375 citations in the ISI Web of Science database from Thompson). This tool takes a list of predetermined "significant" genes as an input. This list may come from one of many methods to identify significant genes (e.g. hierarchical clustering, Significance Analysis of Microarrays (SAM) [136, 137], Gene Set Enrichment Analysis (GSEA) [138, 139], fold change or p-value filtering, or machine learning and classification methods). GOMiner counts the appearance of these genes in association with ontological terms and uses the Fisher's Exact statistical test [140] to estimate a p-value to indicate if "significant" genes are over-represented in a functional category.

In preliminary work targeted at enhancing these types of studies, we have implemented a system that can compare these p-value results among lists of significant genes generated by numerous methods, or generated by heterogeneous data sets. The original GOMiner tool was a standalone Java application that could connect to a local or remote database, but required installation on the user's desktop PC. It also involved a "Windows Explorer"-style hierarchical interface for most use cases. eGOMiner [141] was a web-based system that made use of SVG visualization and interactive navigation to enhance the user experience. The eGOMiner system contained many graph layout problems, including unnecessarily crossed lines, limited ability for functional results comparison, no consideration of the potential for graphs to overwhelm the system as the

Gene Ontology database grew over time, and limited interoperability with other systems in the research workflow. The key improvements in my work on eGOMiner are integration of a much simpler and more effective graph layout system, pre-filtering of terms based on p-value to manage the overall tree size, and definition of a caGrid UML model so that GOMiner functional results can be integrated into other grid-based systems.

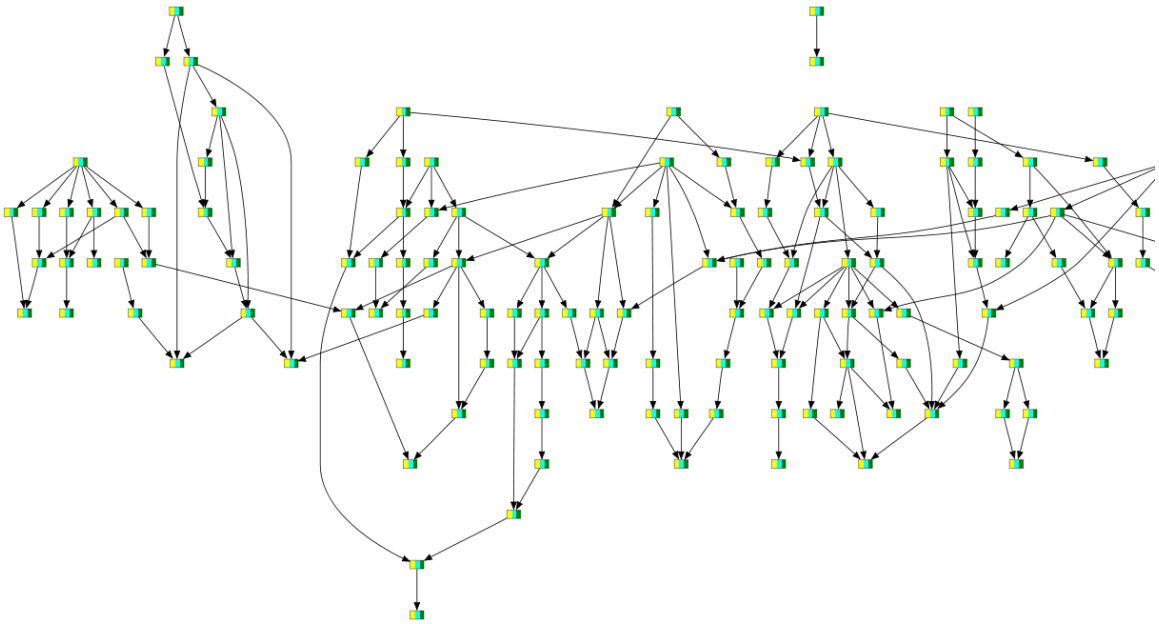


Figure 23: eGOMiner graph visualization of 'biological process' ontology branch of the Gene Ontology. Each node in the Gene Ontology is a biological term (classification of gene function). The nodes nest inside each other as the terms progress from general to specific functional classifications. This visualization uses a combination of GraphViz for generating an appropriate layout and custom code for generating the colored nodes for comparative information. The yellow, aqua, and green sections of each node represent three separate input gene lists under analysis. The opacity of each rectangle represents the p-value result of the Fisher's Exact test of the gene list at that term.

SphingoVisGrid and PathwayVis: Comparing Quantitative Lipid Synthesis Pathway Experiments

The field of metabolic pathways is full of useful examples of how interactions between branching pathways can cause dramatic growth in complexity of analysis and visualization [142]. Certain areas of the biological domains can be simplified so that the representations follow a hierarchy, with limited interconnections among low-level branches. For these domains, two-dimensional graphs can be delivered through web-based tools for interactive exploration on a variety of devices.

One example well-suited to this type of simplification is the Glycosphingolipid synthesis network [143]. Sphingolipids are a component of cell membranes of many species and their intermediates and products regulate diverse cell functions. The metabolic pathway of sphingolipids has thousands of individual components, which are investigated using separation columns and tandem mass spectrometry [144, 145]. These instruments produce high-volume datasets that are challenging to interpret.

Sphingolipids are comprised of a sphingoid base backbone and a variable-length fatty acid chain. The sphingoid base is synthesized *de novo* from serine and an acyl-coenzyme-A and may be converted into ceramides, phosphosphingolipids, glycosphingolipids and other species (for example, phosphocholine for sphingomyelin, and sugars for glycosphingolipids). There is considerable variation in all of the components, including several hundred known variations in glycosphingolipid headgroups, over 70 sphingoid base backbone varieties and dozens of amide-linked fatty acids (see Table 6).

Table 6: Examples of the combinatorial nature of sphingolipid synthesis.

Modifiers	Implemented in SphingoVisGrid	Identified in literature	Theorized based on chemical properties
Head Group	8	150+	400+
Backbone Chain Length	12	9	20+
Backbone Chain Double Bond Variants	2	3	3
Fatty Acids	2	3	3
Total Combinations	384	12,150	72,000

PathwayVis is a web-based and database-driven visualization tool that stores experimental and simulation results and can compare the concentration changes in the context of the pathway (see Figure 24). This system has been used to support studies of the GlycoSphingolipid biosynthesis pathway [146]. This pathway has been partially loaded into the LipidMAPS database [147], and the PubChem database of chemicals [148]. Scalability of this visualization is an extremely important feature of this software, because of the combinatorial nature of the structural components of the chemical species.

Quite a few popular pathway visualization tools have been developed, including CellDesigner [149], CytoScape [106], COPASI [150], and JSim [151]. One system that is tailored for a specific community is GlycoVis [152]. Most of the many tools available are only available for running locally and do not support web-based visualization or grid-enabled APIs. Computer scientists have extended the basic graph visualization work offered by GraphViz using techniques such as 2.5D pathway maps [153] and Focus+Context hierarchical navigation [154].

PathwayVis and SphingoVisGrid are differentiable from previous work in the technical implementation decisions and in the design objectives. The technical decisions enable web-based access and exploration, including through popular mobile devices (see

Figure 26). The design objectives make this system customized specifically for the lipidomics community instead of trying to solve a general problem. Systems that treat all problems as the same tend to compromise the usability and extend the learning curve of the system for non-computational biologists. Technical decisions were made to enhance the usability of SphingoVisGrid, leading to the use web-enabled PHP and SVG programming technologies and support for uploading Excel spreadsheet data already in use by experimental scientists using the comma separated values (CSV) format (see Figure 25). The user interface supports navigation of time series data and will create a pie chart graphic to support comparisons between similar chemical species in the pathway.

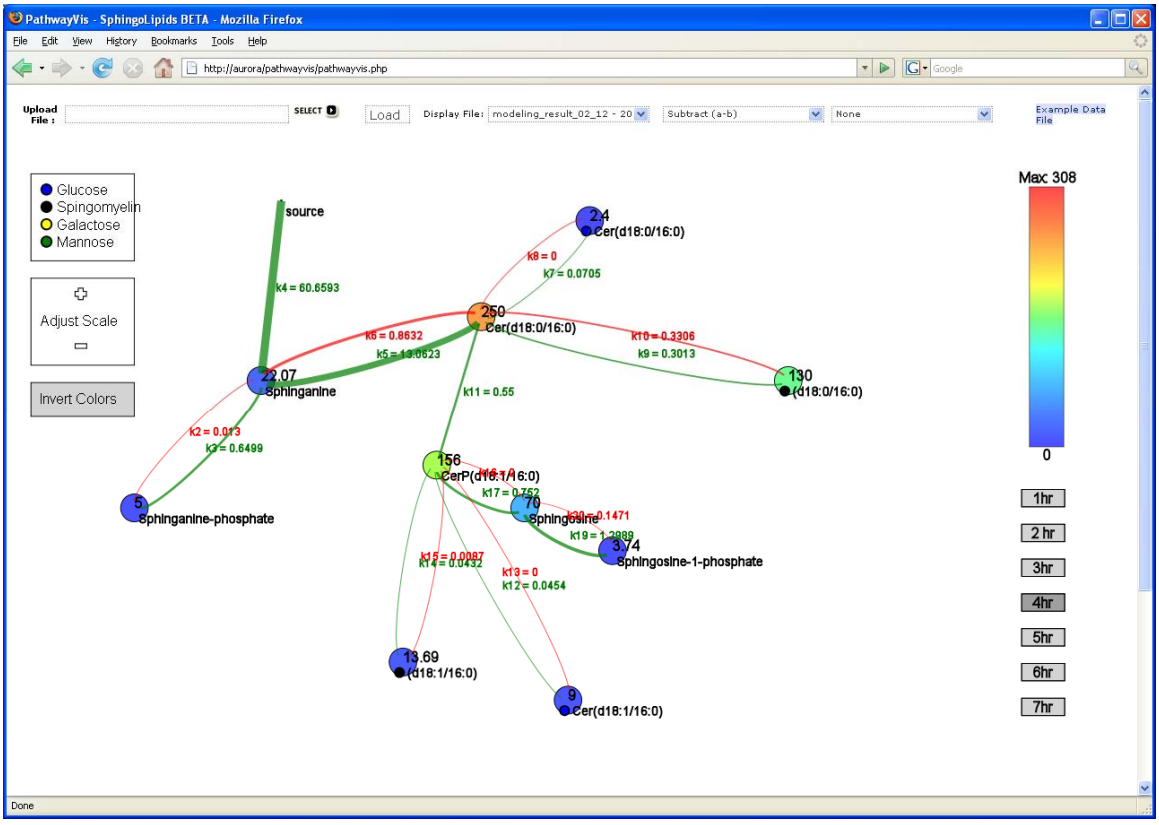


Figure 24: PathwayVis screen shot. This SVG rendering of a partial sphingolipid biosynthesis pathway displays the k-values predicted by a parameter estimation algorithm on the edges and the difference between the simulation predictions and the empirical dataset used for training on the nodes.

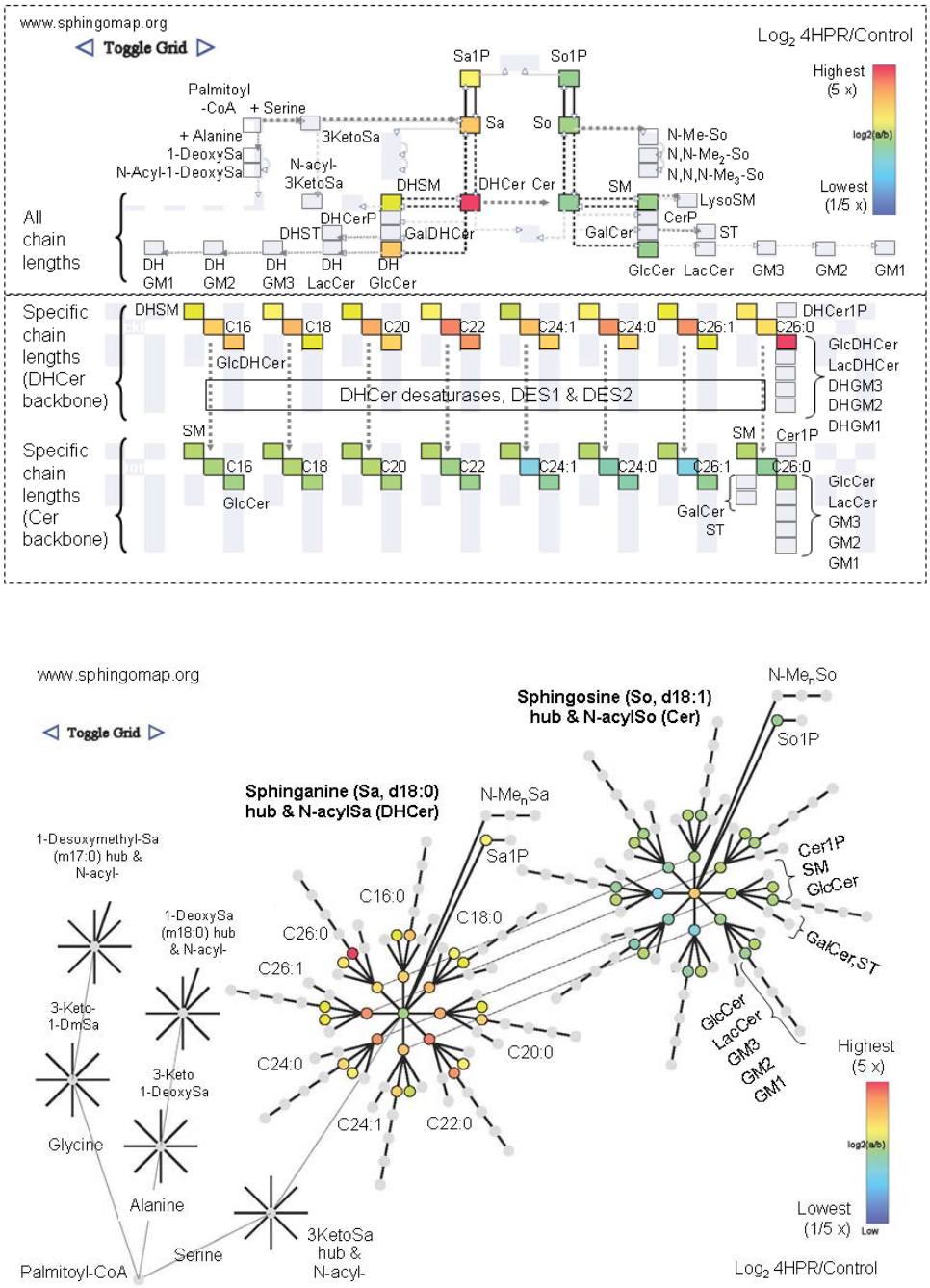
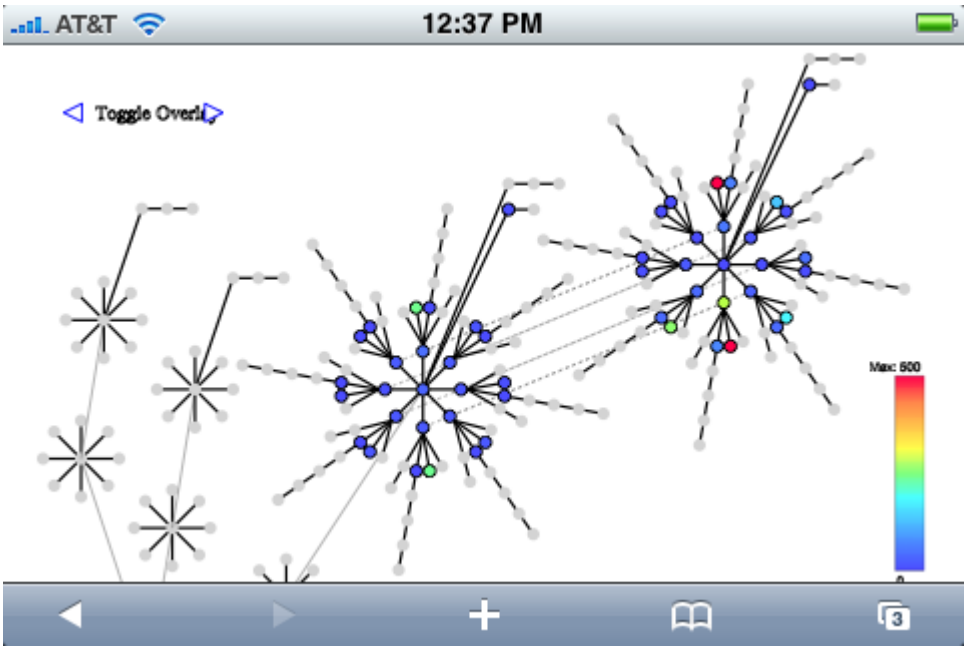


Figure 25: Annotated SphingoVisGrid screen shot. This SVG rendering contains two views of the same data measurements. In the first view (top), the data is formatted similarly to the data input file. Both the data input file and the grid visualization were organized conceptually by the biological domain scientists. The second view (bottom) displays the data in a branching radial view that is more indicative of the biochemical pathway. Both views have been annotated by a domain scientist (Dr. Al Merrill) to reflect details of the molecules under study. The system is flexible enough to support branching pathways of low or high complexity. These figures are used with permission from [155].

A)



B)

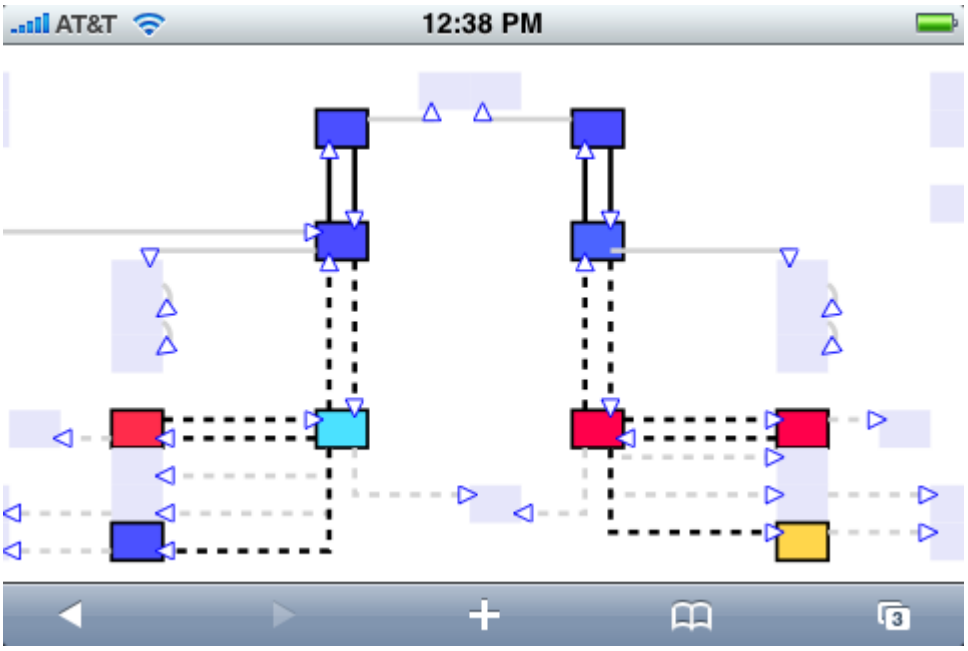


Figure 26: SphingoVisGrid on the iPhone mobile device from Apple. One of the primary motivations for using SVG in the SimpleVisGrid visualizations is that it is becoming widely supported for mobile devices. The graphics are represented mathematically, so panning and zooming are simple to implement and always produce crisp images, which is crucial for small screen space devices. A) The nodal view. B) The same experiment in the grid view, zoomed in (~4x) on a complex part of the pathway using the touch controls. The iPhone platform currently does not support Adobe Flash applications because of performance problems.

Visualizing Correlations: Feature Landscapes and Correlation Heatmaps

Another important aspect of comparing widely varying and multi-scale biological data is the ability to detect similarity and differences among massively high-dimensional datasets. Feature landscapes are an example of using the SVG format to depict 2.5D visualizations. Although the peaks and the background appear to have depth, they are positioned on the graphic using pre-built symbols. Each peak symbol contains custom attributes providing an ID for the feature and the exact number of times it appears in the provided data, as this information could never be recovered from the X,Y location of the peak in the graphics file. I have used feature landscapes in my research to see if tuning certain parameters in feature selection algorithms significantly affects the results. I am developing an algorithm for summarizing the similarity between the feature lists for a future publication (see chapter 5).

For microarray studies, it may be useful to compare the analysis results of many different teams to determine how much agreement there was among the teams on the most important features of the data. This was an important contribution to the FDA MAQC Phase II project because over 30 teams attempted to analyze the same 13 endpoints of the six provided datasets. Gene landscapes are a rich source of information because they combine a sense of the scale of the problem (the amount of gray space and the narrowness of the feature peaks) with a measure of frequency of selection of certain features in the large solution space.

Many studies (e.g. Genome-Wide Association Studies (GWAS)), attempt to correlate disease states with genomic variations using datasets of large sample size (thousands of genotype study results). These correlations are not necessarily the endpoint of the study, but must be collected in a large database for searching and querying operations in order to test hypotheses. For this reason, results of these correlations are

stored in a BioPNG format and provide query operations to quickly retrieve the images into memory and retrieve a row, column, or single pixel from the image.

High-dimensional, high-throughput data acquisition methods have a tendency to suffer from a phenomenon called “batch effect.” This is an overall bias applied to the data based on the period of time that the data was acquired. Batch effects may be caused by environmental factors, human error, changes in equipment or changes in lab protocols. It is very important that these problems are corrected before data analysis models are built, because the bias can become the dominant feature of the data.

Our meta-data correlation visualization can be used to detect batch effect, but can also be used to evaluate experimental design problems, such as sudden changes in chip quality, inappropriate randomization procedures, and relationships between clinical factors that might be useful to exploit when building a data analysis model. These uses are discussed fur-ther in chapter 5.

caBIG Certification Preparation of SimpleVisGrid

The reason SimpleVisGrid is simple is because only two input formats are accepted and only two output formats are provided (see Figure 27). SimpleVisGrid services are based on a UML Model constructed in Enterprise Architect (see Figure 28). Common Data Elements (CDEs) have been identified and semantically annotated using the Semantic Integration Workbench (SIW) provided by caBIG. Where necessary, new concepts have been identified and defined for insertion into the Enterprise Vocabulary Services (EVS) ontology for cancer research. These practices mean that we can build a submission package for Silver-level certification of SimpleVisGrid.

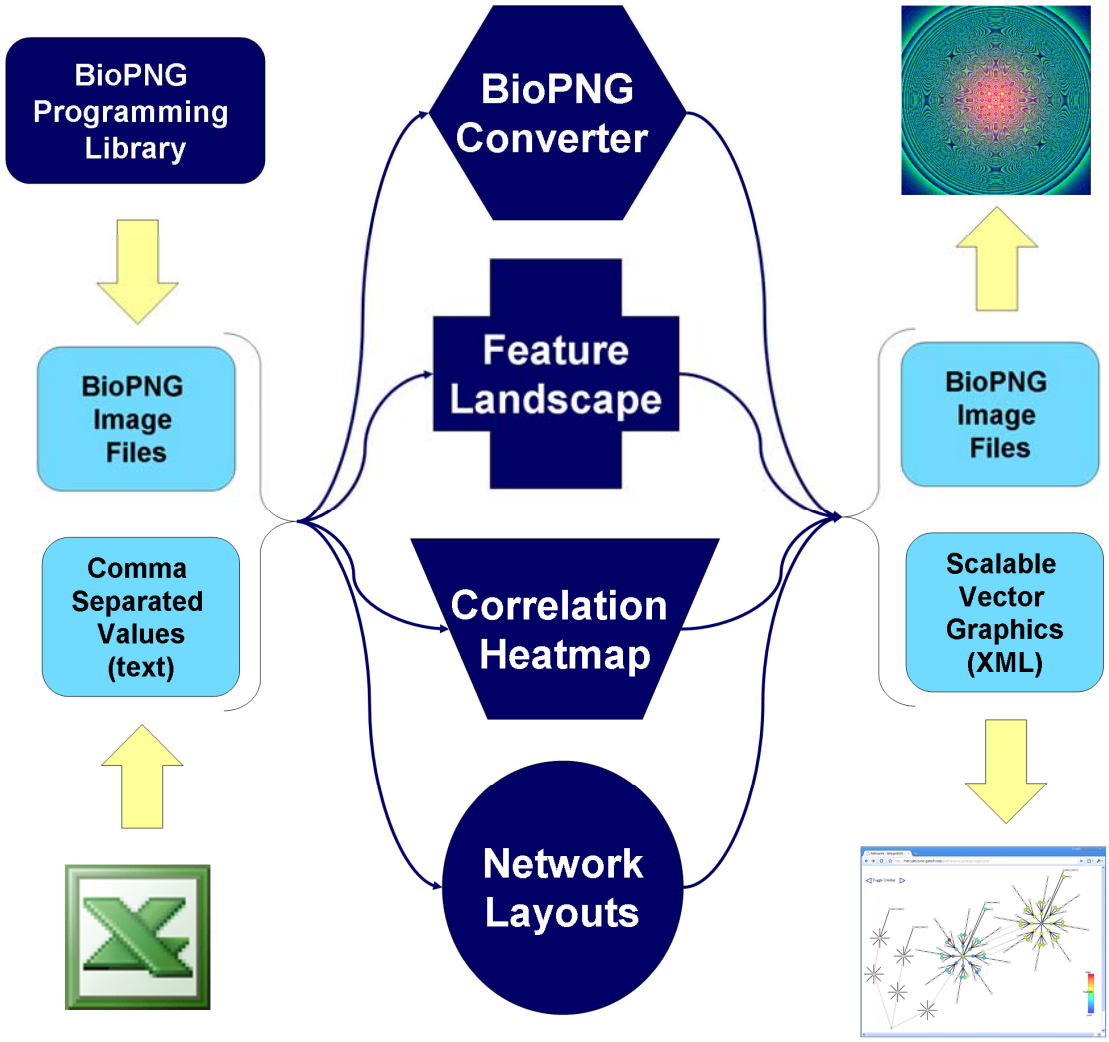


Figure 27: Schematic of the grid services making up SimpleVisGrid 1.0.

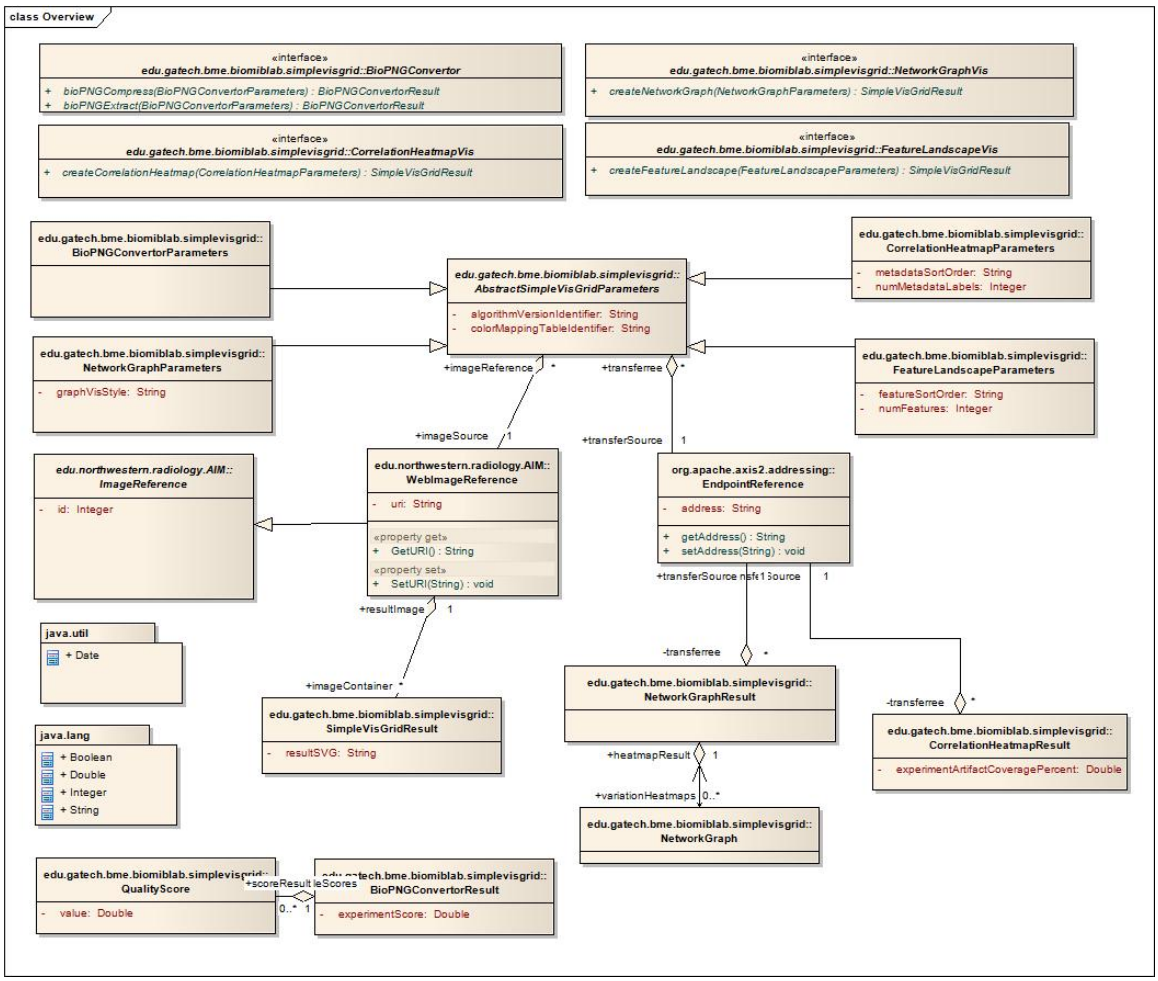


Figure 28: Semantically annotated SimpleVisGrid UML Model describing the input and output objects of the SimpleVisGrid services.

CHAPTER 5

FDA MAQC PHASE II CASE STUDY

Participation in the United States Food and Drug Administration (FDA) Microarray Quality Control (MAQC) Project provided an opportunity for me to apply the expertise I acquired in developing caCORRECT and ArrayWiki to investigate the impact of additional microarray experimental quality factors on clinically-relevant problems. My work was included in three papers submitted for publication in the MAQC special issue of *Nature Biotechnology*. This work is currently under review, so in this chapter I provide a brief overview of my contributions.

Background on MAQC

Microarray gene expression data have been submitted by pharmaceutical companies to the FDA for a number of years to support claims about the safety and efficacy of Investigational New Drugs (INDs) and New Drug Applications (NDAs). It is important that the FDA conduct independent studies about the microarray technology before accepting this evidence in the place of more traditional approaches. Since the FDA is the gatekeeper for technologies that are in use in real medical practice, the impact of this work on translational biomedical informatics is large.

The reliability of microarrays as a reproducible measurement technology has been brought into question by a number of researchers [156-162]. The MAQC Phase I Project demonstrated the technical reliability of microarray technology in detecting differential gene expression and was published in 2006 [42, 163, 164]. The FDA MAQC II Project answers questions regarding the reliability of the data analysis around the technology for clinical applications. Many different approaches to solving the problem of building predictive models were encouraged in order to explore the territory and ensure that the guidelines developed cover the range protocols in common use today.

Among the approaches developed by the 36 teams, the steering committee found the K-nearest neighbor (KNN) method proved to be one of the simplest and most robust classifiers. Despite this, large variations in prediction performance still occurred among the KNN models. The Georgia Tech team volunteered as an unbiased outside observer to investigate why this variation occurred. We designed a significant systematic study inside the MAQC project that would serve as a mini-MAQC. I contributed to the FDA MAQC Phase II effort in three ways: (1) I collaborated with John Phan in the development of models to be submitted for the Georgia Tech team, (2) I developed testing harness software to fully exercise the K Nearest Neighbors (KNN) classifier across a wide variety of parameters, and (3) I developed visualizations to help understand the meta-data

Development of Models for MAQC-II

caCORRECT was featured in the construction of models from Georgia Tech for testing the predictive power of gene expression results on three Affymetrix microarray platforms. Other methods employed included a genetic algorithm for selecting small combinations of genes and mining of the Gene Ontology for genes that were functionally related to the clinical scenario. Public results of this effort will be published on ArrayWiki along with accompanying source experimental data once the consortium has published its results and the data is no longer confidential.

Investigation of Differential Performance of KNN

The KNN classifier uses simple distance metrics between all points in a set of training data to build a set of size K neighbors. The neighbors then vote to choose the class of a new testing point based on their own class labels. Despite the simplicity of the KNN method, results from the MAQC-II Project suggest that there are factors within the population of KNN models that cause large variations among prediction performances by different teams. Given the fact that KNN is a common method used in data analysis

studies, it is an ideal proof-of-concept classifier for us to gain insight about reproducibility and reliability in microarray-based predictive model development.

We designed a combinatorial study of 463,320 KNN models by varying nine parameters for each of the ten endpoints from three clinical datasets: breast cancer [165], multiple myeloma [166], and neuroblastoma [167]. The parameters varied were (see Figure 29):

- K (number of neighbors) on a range from 1 to 30.
- The metric used to calculate distance (euclidean distance, cosine distance, or city-block distance)
- The method of feature selection (SAM, P-value, or fold change)
- The number of features selected (5:5:200)
- The method used to calculate voting among the neighbors (equal-weight, prevalence-weight, and distance-weight)

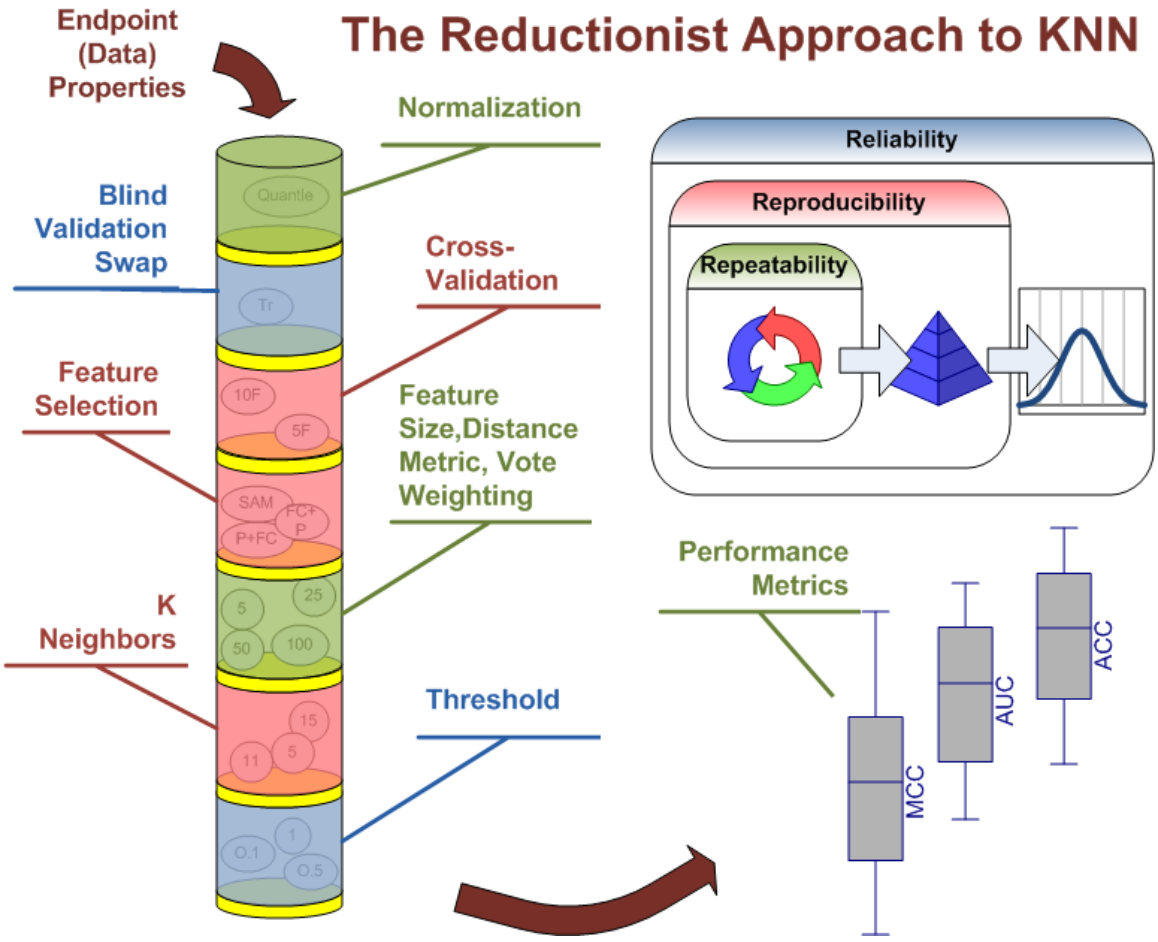


Figure 29: Workflow of the KNN investigation. Our system design is similar to the reductionist approach employed by biologists to extract components of the system that are related to each other by fundamental properties. The metaphor of the column on the left is that of a liquid (or affinity) chromatography column. Four datasets were passed through the same combinatorial approach, where a parameter was cycled through 3-30 possible values to create a comprehensive results dataset. That dataset was mined using ANOVA to determine the relative contribution of each parameter type to the performance variation. Each parameter was also classified according to its impact on reproducibility of results on the same data (measured during cross-validation) and on external data (measured using the blind validation dataset provided by the FDA).

Identification of Modeling Factors Affecting Performance.

Analysis of Variation (ANOVA) was used after separating models by each factor studied to determine the relative effects of modeling factors on performance. Each modeling factor is selected in turn and a one-versus-rest comparison is used to measure variance. The results of a preliminary ANOVA are shown in Table 7. These results show that the impact of KNN-specific parameters is dwarfed by the impact of innate data properties. Thus, it seems that in looking to improve performance of KNN models, addressing problems in the data can contribute the most to improvement.

Table 7: Sources of variation (ANOVA) in external validation performance across six FDA clinical scenarios.

Data set	Perf. Metric	Sample Method	Rank Method	Feature Size	K	Vote Method
53999.7	1832.7	205.9	3.1	195.7	192.3	151

Analysis of Dataset Properties.

I employed a new quality control technique in the process of investigating why one participant in the MAQC project reported significantly better results than others using similar methodologies. We hypothesized that the source of this performance advantage was a method of dropping samples from their analysis that might contain confounding data. We compared a chip-to-chip correlation plot with various technical features of the data for four datasets to verify this hypothesis (see Error! Reference source not found. and Figure 30). The resulting visualization was found to be even more useful for identifying data batch effects than the gel plots developed for ArrayWiki.

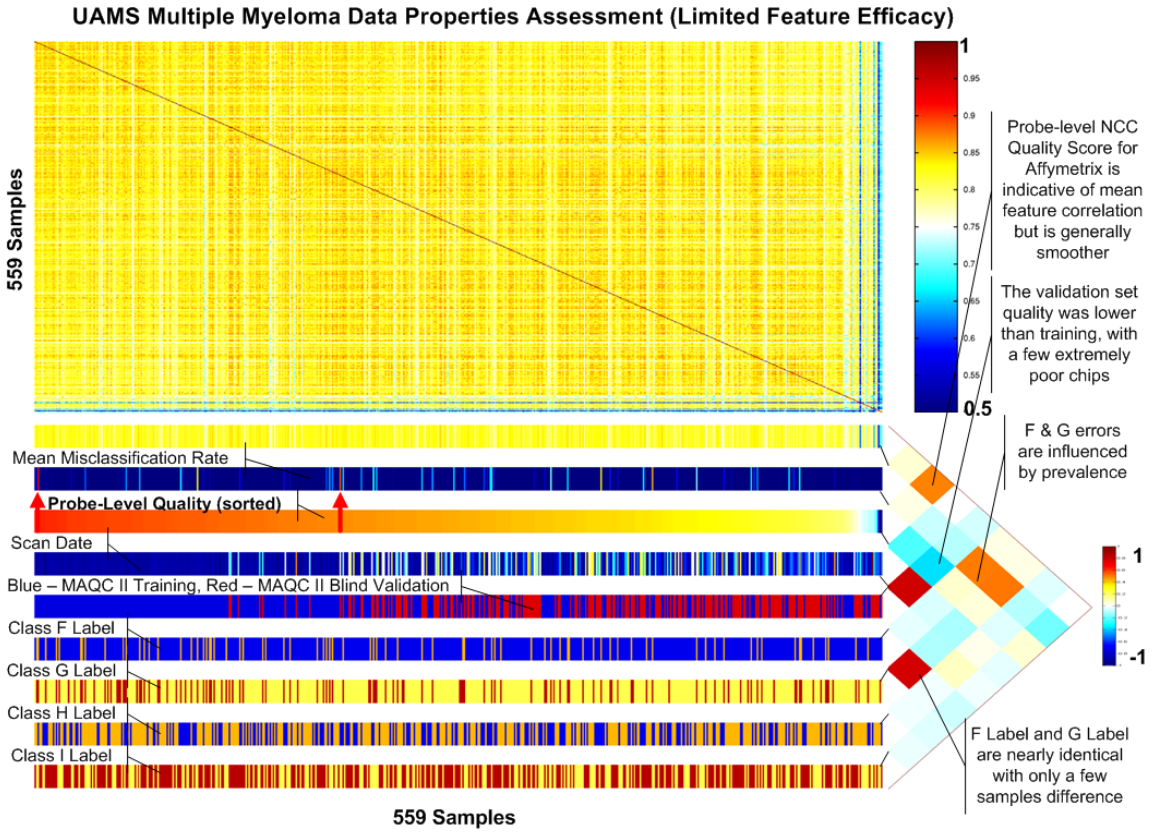


Figure 30: Chip-to-Chip correlation plot for FDA MAQC-II multiple myeloma dataset. The colored bars along the bottom allow someone to draw a variety of conclusions. First, you can compare the selection of a training set to the validation set. The method of sample collection and assignment to training/validation is non-random (as observed by the red diamond where Scan Date meets the training/validation labels. This will make a dataset more prone to batch effect bias. This is the largest dataset available from MAQC and exhibits no batch effect like that found in other datasets. A few chips have very poor scan quality, but they do not appear to have a high misclassification rate. The two samples indicated by red arrows might indicate borderline or incorrect clinical labels. The F clinical endpoint label is a complete subset of the G clinical endpoint label. This indicates that a joint classifier would work better than two independent classifiers.

Visualizations for Understanding MAQC Meta-data

In addition to working on a project specific to our Georgia Tech team, I contributed some visualization expertise to support the claims of the MAQC main paper led by Dr. Leming Shi. My primary role in this effort was to compare the lists of features selected by all 36 teams to determine how much agreement there was among the teams. The agreement was found to vary considerably depending on the difficulty of the clinical scenario and the quality of the data. I used feature landscapes to visualize the amount of agreement, both on the original training data and in comparison to the building of models on the blind validation data during the "swap" analysis. I developed two metrics to allow for summarizing the agreement among feature lists into a single number. One metric works best for unranked lists of the type submitted by the teams for their models. In this case I used the Fisher's exact statistical test to generate a p-value of the likelihood that the two lists were generated by random chance. The second metric is based on a novel algorithm that takes into account the ranking of two lists, and was used to compare the lists produced by the three ranking methods for the KNN study. Both metrics work well when the lists have different numbers of elements, unlike many existing methods such as Borges Count or Canberra Score.

Feature Landscapes

The feature landscape is a 2D histogram where estimates of molecular concentration or any other biological feature that might make a useful biomarker are arrayed using any sort order. The histogram is presented in 2.5D to give a more dramatic presentation, and to guide the eye at quickly recognizing pairs of peaks when comparing two landscapes. Peak heights provide for quicker recognition than using color only. Figure 31 shows landscapes generated for the KNN study of three simple feature ranking methods: Significance Analysis of Microarrays (SAM), fold change ranking after P-value filtering (FC+P), and P-value ranking after fold change filtering (P+FC).

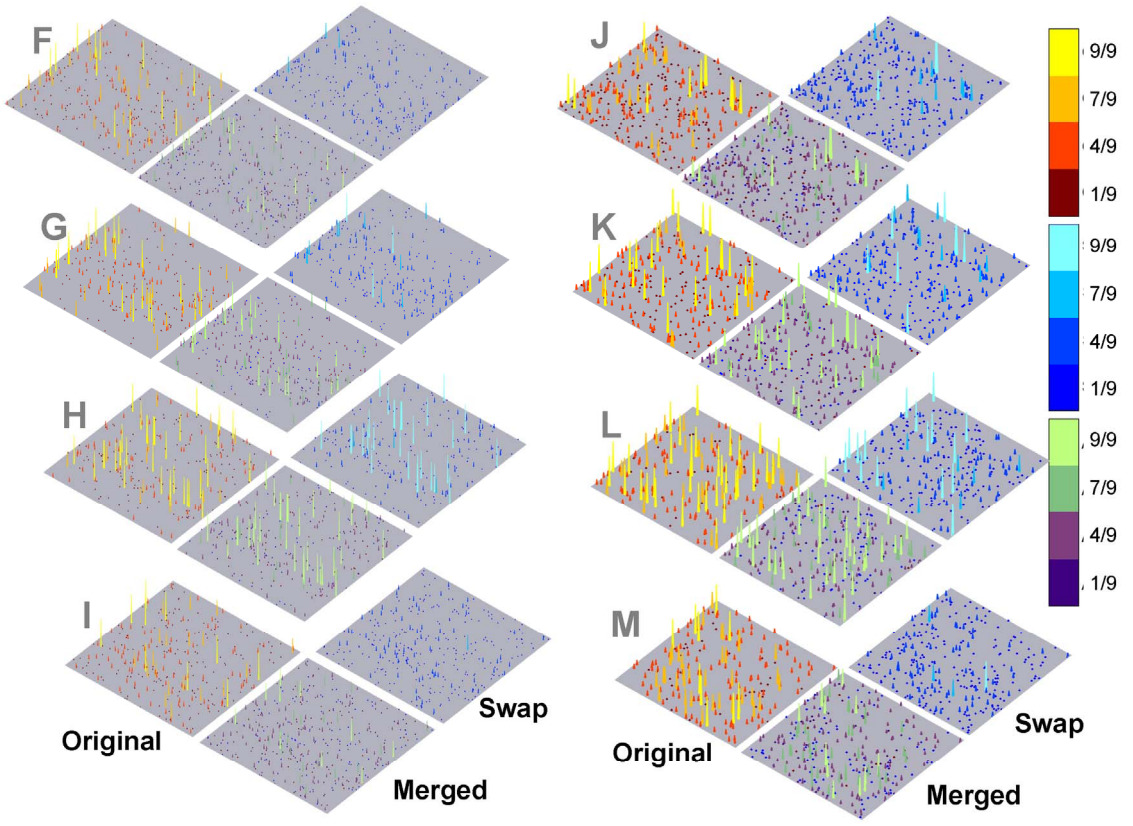


Figure 31: Feature landscapes comparing features lists generated by KNN study. Nine feature ranking methods were compared while testing the KNN "full monty" systematic study. The three fundamental types were SAM, fold change, and p-value. For fold change and p-value, there were a few minor parameters that created four unique methods for each fundamental type. The landscapes with the hot colormap show agreement between the nine methods using training data and the cold colormap shows agreement on validation data. The merged colormap (green and purple) shows agreement among feature selection methods between training and validation datasets. Agreement is higher for training in general because it has a larger sample size. Agreement for I and M is small because the class labels were randomly assigned as a negative control. The design of the microarray platform for endpoints FGHI is the most likely reason for the bias toward features in the middle band of the gene list.

Development of Distance Metrics for Feature Lists

The feature landscape visualization gives an impression of overall agreement of feature ranking methods and can confirm expectations about difficulty or point to experiment design problems. However, it is also useful to generate a single number that measures the similarity between two lists. In working on the MAQC-II project, I developed a list of two shortcomings of microarray classification studies that were not addressed in the literature, and which could be improved:

1. Reporting of "unranked" gene lists. Ranked gene lists necessary for solid comparisons between competing gene lists and effective merging of lists from different studies. Unranked gene lists are selected using an arbitrary cut-off based on various performance metrics. Even when the cut-off is reported, it does not aid in further trimming the gene list or prioritizing the gene list for external validation. The best ranking beyond a discrete numbered list is a true confidence score reported by the ranking algorithm.

2. Lack of a metric for comparing similarity between two feature ranking results. This metric might be useful for comparing to "knowledge" such as previously validated biomarkers or knowledge gleaned from literature or stored in a database such as the Gene Ontology (GO). Existing methods, such as Fisher's Exact, assume unranked lists, but a more precise measure can be obtained when using ranked lists.

I addressed these issues by adapting the existing Fisher's Exact statistical test as a distance metric for unranked feature lists and by developing a new distance metric for ranked feature lists. These methods can be extended to problems beyond biomarker identification using microarrays. There are many examples of problems in biology where ranked and unranked asymmetric sets are compared and many of these scenarios could make use of a distance metric for quantifiable comparisons.

Normalized List Similarity Distance for Ranked Lists

R(x,X) is the rank operator, which returns the integer rank > 1 of the x element of the ranked list X, or the confidence value > 1 of that element. L(X) is the length of the ranked list X.

$$A_{spread} = \sum_{i}^{L(A)} \frac{1}{R(\alpha_i, A)}$$

$$LSD(A,B) = \sum_{i}^{L(A)} \begin{matrix} \alpha_i \in B \Rightarrow \left| \left(\frac{1}{R(\alpha_i, A)} \right) - \left(\frac{1}{R(\alpha_i, B)} \right) \right| \\ \alpha_i \notin B \Rightarrow \frac{1}{R(\alpha_i, A)} + L(A) \times A_{spread} \end{matrix} + \sum_{i}^{L(B)} \begin{matrix} \beta_i \in A \Rightarrow \left| \left(\frac{1}{R(\beta_i, B)} \right) - \left(\frac{1}{R(\beta_i, A)} \right) \right| + \left| L(A) - L(B) \right| \\ \beta_i \notin A \Rightarrow \frac{1}{R(\beta_i, B)} + L(B) \times B_{spread} \end{matrix}$$

The normalized List Similarity Distance is simply the calculated value divided by the value of the worst case scenario (i.e. a complete list mismatch).

$$normLSD(A,B) = \frac{LSD(A,B)}{A_{spread} \times (L(A)+1) + B_{spread} \times (L(B)+1)}$$

A calculation must meet the following criteria for all x, y, z in X to be considered a distance metric:

1. d(x, y) ≥ 0, (a.k.a. non-negativity)
2. d(x, y) = 0 if and only if x = y, (a.k.a. identity of indiscernibles)
3. d(x, y) = d(y, x), (a.k.a. symmetry)
4. d(x, z) ≤ d(x, y) + d(y, z), (a.k.a. triangle inequality).

Of these, the first three can be generally satisfied by inspection. The fourth requirement is more difficult to prove. It was ultimately proven for lists up to six members in size by exhaustive search (see Figure 32).

Study of Feature Selection Stability Among Common Ranking Methods

In the objectives of the MAQC-II project, reproducibility of features among differing data analysis protocols (DAPs) was not emphasized as the most important quality measure. However, for clinical applications of microarray classifiers, including selection of candidate patients for personalized treatment with new therapies, a fixed and validated set of biomarkers must be arrived at in some way. So, as a way of looking at the end goal of many readers of the MAQC-II results, we consider the reproducibility of feature lists of proven biological significance (i.e. that will hold up to validation using methods more sensitive than microarrays).

Accurate measurement of these two qualities of a classifier is an open question. Our metric emphasizes the rank order of each feature list, the size of intersection between two lists, and the difference between the two lists. Figure 33 shows the results for 10 endpoints varying the number of features from 25 to 125 in steps of 5. A distance of zero (light blue) indicates perfect matches between gene lists. A distance of 1 (dark blue) indicates complete disagreement between lists. Of the three feature selection methods we evaluated, we found that ranking by fold change with a p-value threshold of 0.05 (FC+P) provided more robust gene lists in terms of genes being likely to appear in similar ranks on lists generated across the 50 cross-validation gene lists. This is particularly apparent for the mid-range difficulty endpoints: D, F, G, J, and K.

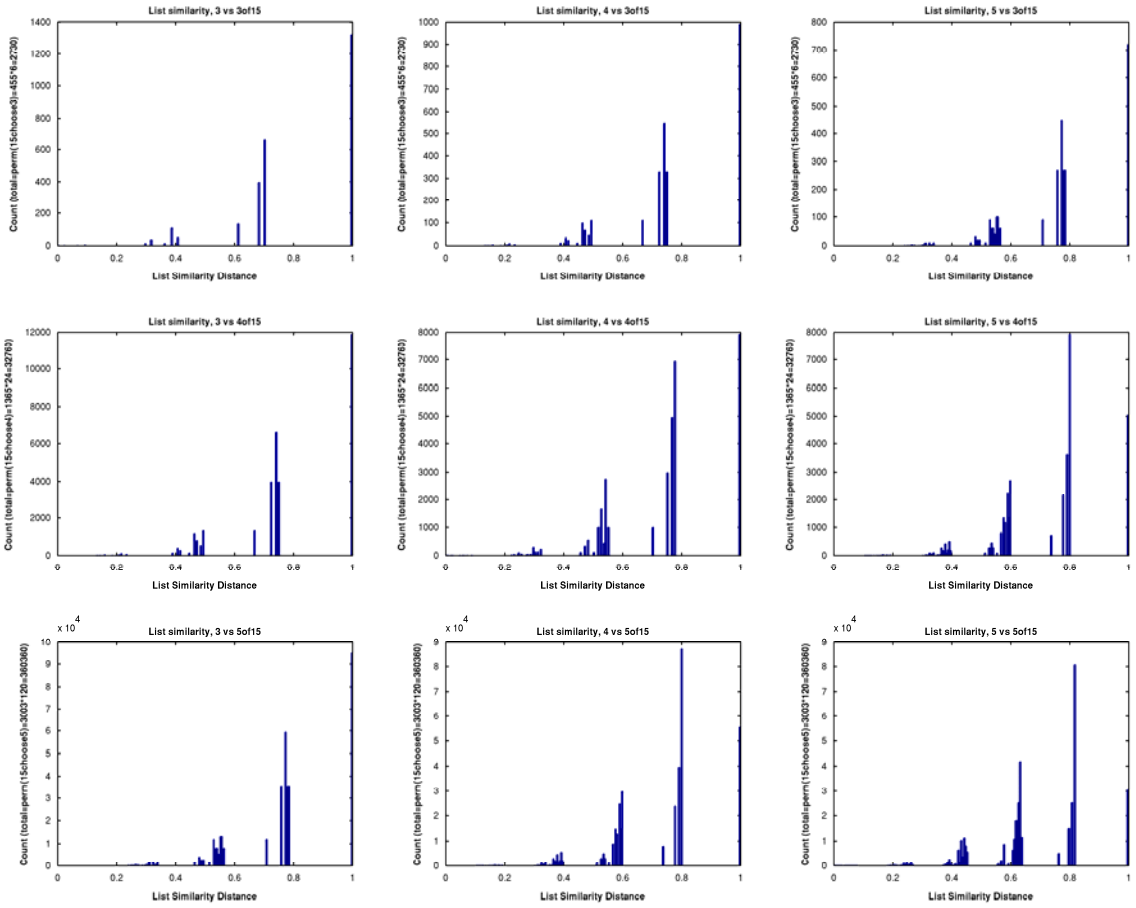


Figure 32: Histograms of complete scoring results of all possible permutations of lists of varying sizes. This series of histograms showing the distribution of values for the list similarity distance metric with lists of varying sizes and a total pool of features of 15. The data in the histogram are the results of scoring every possible permutation of lists from the 15 features exactly once. The penalty associated with mismatches is the reason for the separation between the segments of the histogram. This penalty varies based on relative lengths of the two lists. The overall trend is for very few of the possible permutations to score well and the majority to score poorly, which is the reason for the very high significance threshold.

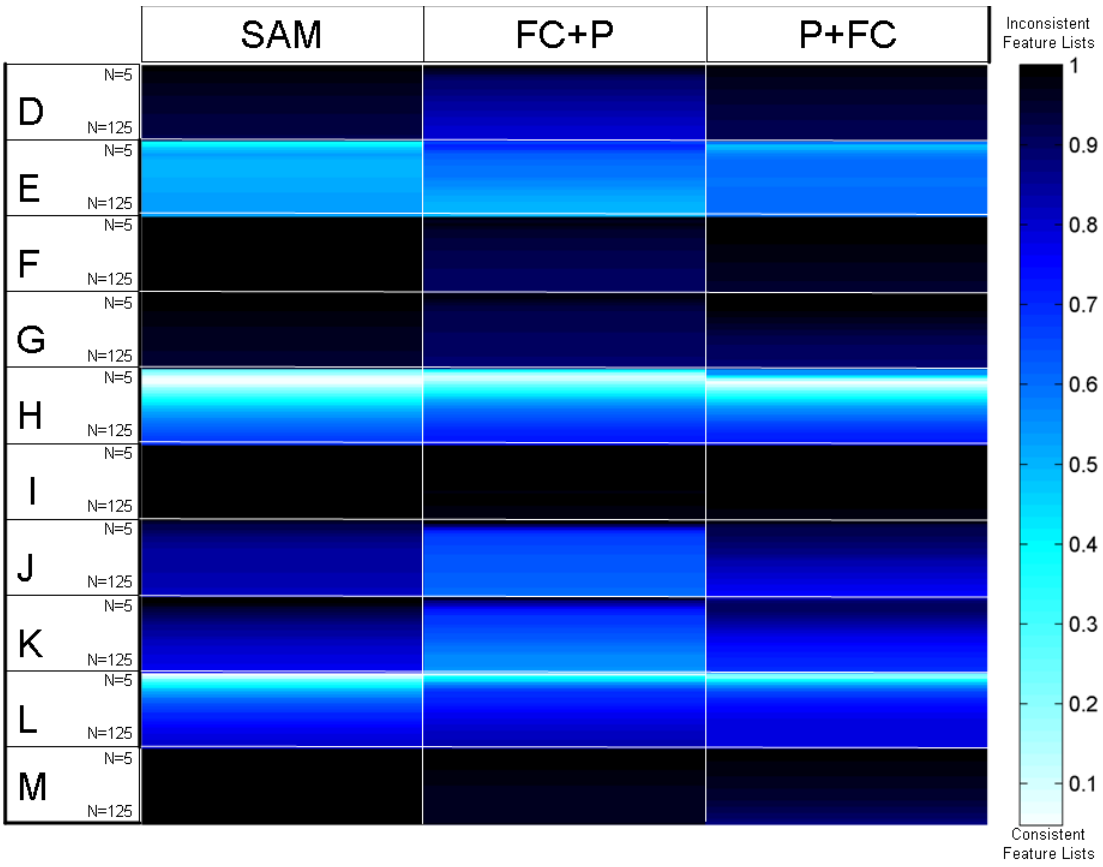


Figure 33: Feature set concordance among cross validation sets for each feature selection method and endpoint. Black indicates a large average distance, or difference, among feature sets from different cross validation folds or iterations. White indicates a smaller distance, or more similar feature sets. Within each endpoint, feature size increases from top to bottom. As feature size increases, concordance of lists tends to decrease. The three ranking methods compared here are Significance Analysis of Microarrays (SAM), fold change ranking after P-value filtering (FC+P), and P-value ranking after fold change filtering (P+FC).

CHAPTER 6

CANCER BIOMEDICAL INFORMATICS GRID (caBIG) CERTIFICATION

The variety of biomedical data and tools has exploded in recent times. However, many such research tools will never be used by the larger user community because they were not interoperable with variety of platforms. The cancer Biomedical Informatics Grid (caBIG) has created a collaborative grid network addressing this issue through a semantic model-driven framework. In this chapter, I describe the path taken to re-design the microarray data quality control package, caCORRECT, to make it available to a wider clinical community through caBIG Silver-level compatibility review. Key software elements and functionality were identified and annotated following grid standards. Clinicians trained to use these tools may be confident that a strong community of developers is committed to delivering interoperable tools of the highest quality.

Translational Biomedical Informatics (TBMI) will allow for much greater efficiency in sharing patient medical records and accumulating de-identified research data where allowed for by federal regulations. There are still many social as well as technical hurdles to overcome on this goal, but scientists and national leaders alike agree that this is an important mission for resolving the many problems related to the rising expenses of health care research, translation, and delivery.

The establishment of a central ontology for cancer research called Enterprise Vocabulary Services (EVS) led to the merging of many of the leading medical ontologies. The Cancer Data Standards Repository (caDSR) requires that all identifiers used by grid-enabled, caBIG-compatible software be mapped to their component concepts using codes found in EVS. This laborious process will go a long way in making the function of software systems in medicine more understandable to doctors. The deployment of simple services (and accompanying new terminology) to caBIG is part of

the third objective of this dissertation for this reason. This work has been submitted to the American Medical Informatics Association (AMIA) conference and is under review.

Cancer Biomedical Informatics Grid (caBIG)

The caBIG project was launched in July 2003 at the National Cancer Institute (NCI) with the goal of creating the "World Wide Web of Cancer Research" [168]. caBIG is part of a national initiative to improve the health care information technology infrastructure to streamline the path from clinical data collection about patients to computational mining of the data to identify molecular profiles for personalized treatment [72]. It has been described in Nature as a "gradual revolution in working practice. [169]" Even this characterization is hopeful, as day-to-day operational challenges in the clinic present unfamiliar territory for the spread of computational research. The caBIG organizational structure is divided into working groups, each focused on different aspects of the interoperability problem specific to cancer research. Two overarching work groups, "Architecture" and "Vocabularies and Common Data Elements," have begun to produce some very detailed specifications that will be extremely useful for guiding the design of interoperable bioinformatics tools [170, 171].

The primary benefit of caBIG is that it provides a common platform from which to launch more ambitious integrated solutions. Bioinformatics laboratories can now adopt the basic infrastructure, databases, and functionality provided by caBIG working groups to fill gaps in knowledge and tedious information management responsibilities. This allows researchers to focus on their areas of expertise while still offering a complete solution to clinical problems such as tissue bank and equipment inventory, clinical trials management, and vocabulary services.

caBIG project's aim has always been to raise the interoperability of tools and data relevant to bioinformaticians, biological and bio-molecular researchers, nanotechnologists and clinicians working to fight cancer [71, 168]. The need for research

tool and data interoperability is increasingly recognized as part of the critical path for advancing medical research. Studies show that data sharing between biomedical researchers is on the rise, with possible causes being perceived increased impact of the scientific work as measured by citation counts [2, 29]. Researchers can share raw experimental data [172, 173] as well as their developed tools and results [25, 174] by porting their own code or by adopting and populating existing caBIG tools and databases. Physicians can benefit from clinical management tools and data integrated into the network [175, 176].

The benefits of involvement in the Grid extend to developers. The open source structure makes integrating with existing tools easier. The development framework also allows for some automatic code generation and the involved user community makes deployment of new algorithms much faster and more rewarding. Despite these advantages, very few independent groups have submitted tools for caBIG Silver-level review. In fact, we were told by our mentors that we are the first development team to reach this milestone without funding support from caBIG.

Most cancer research data is still isolated into the local control of groups responsible for creating and storing it. Making the data publicly available requires a fair amount of labor both in interpretation and arrangement before it is useful among collaborating groups. This phenomenon has been called an Information Tower of Babel and many current health care problems can be traced back to this root cause [9]. One trend is to amass data from many sources and put them into a single system for storage, analysis, and distribution. This trend most often results in data stores that are stale and difficult to maintain [20, 21]. Development emphasizing interoperability between disparate groups allows myriad systems to connect in a manner similar to the World Wide Web, allowing the network to grow in a federated manner. Efforts from academic, industry, and government entities can integrate their software and data all the while maintaining local control over both. Data pertaining to tissue samples, genetics,

treatments, and clinical trials can be made interoperable in a participatory process utilizing a layered set of standardized vocabularies. Tools responsible for housing and facilitating access to as well as analyzing such data are also good components of a comprehensive biomedical grid.

In addition to these shared tools and architectural guidance, caBIG is a test bed for development of standard tools that can be understood, verified, and used by clinicians around the world. In its application workspaces, caBIG provides oversight and communication channels between researchers at cancer centers around the United States to improve the level of trust in tissue banking, clinical trials, and molecular data mining tools. The caBIG steering committee also regularly organizes surveys to gather feedback from clinical users and provide direction for the entire effort.

There is an important potential for added value in the form of automated knowledge discovery workflows [177]. Just as news readers on the Internet can now set up alerts and crawlers that collect information as it appears on certain web sites, researchers can build larger functional blocks of research data processing that can perform a data analysis protocol on new data as it appears on the Grid. This potential for greater automation of data analysis is one of the key drivers of adoption of the Grid architecture. Greater automation should reduce data analysis errors and address the problems of "stale data" as well as "stale algorithms" which result from the need for significant human intervention in the process of upgrading computational resources.

The idea of a national bioinformatics grid has also been proposed in the UK [178] and in France [179]. In the United States, the Virtual Observatory [180] and the Cooperative Human Linkage Center [181] are examples of great successes of grid-based science. caBIG may be the largest and most ambitious grid project being undertaken to date given the combination of its domain of focus, variety of data relevant types, semantic interoperability goals, and participatory nature. Already, projects are being initiated by the National Institutes of Health (NIH) for a Cardio-vascular Research Grid

(CVRG), building on some of the infrastructure developed by caBIG and deploying it to another large research community.

Semantic Annotation for Interoperability

Much of existing computing data is only available on systems which have no account of the meaning of the data. Search engines have alleviated the problem by grouping available data with human-understandable terms, but machines are ill-equipped to handle even elementary inquiries across different data sources. The effort to semantically annotate information aims to allow meta-data to provide meaning which can be utilized directly by computer systems (see Figure 34). Key use cases motivating the semantic annotation methods for caGrid include the discovery of tools and data, large scale data analysis, and workflow construction.

Existing caGrid projects, such as caGridPortal, provide for the advertisement and discovery of grid services themselves. Maintainer institutions and available service operations are described by all open services to a centralized index server. This information is then, for example, available on a portal webpage for discovery of open services by anyone. The caGridPortal provides a web-based general search solution for grid services and participating institutions based both by name, location, and kind. Search by institute, domain model, objects using certain concepts, and services using objects related to certain concepts are some possible parameters of search allowed by the portal's indexing methods.

Large-scale data analysis is another critical use case motivating the semantic annotation of data. Integrating different data types from heterogeneous services requires the meaning of data to be interpreted; and so metadata representing meaning allows more automated processing of different data types.

Research facilitation by means of a grid service can benefit from semantic annotation in the linking of research and activity data across multiple sources,

incorporating multiple tools, data objects, and data sources. Grid Services developed by independent groups must be made to agree on the meaning of the data being passed between them. In some cases, adapter services will have to be designed to transform data between simpler and more complex structures based on the requirements of the next Grid Service in the workflow. A well-annotated model will make this task simpler and reduce the chance of error.

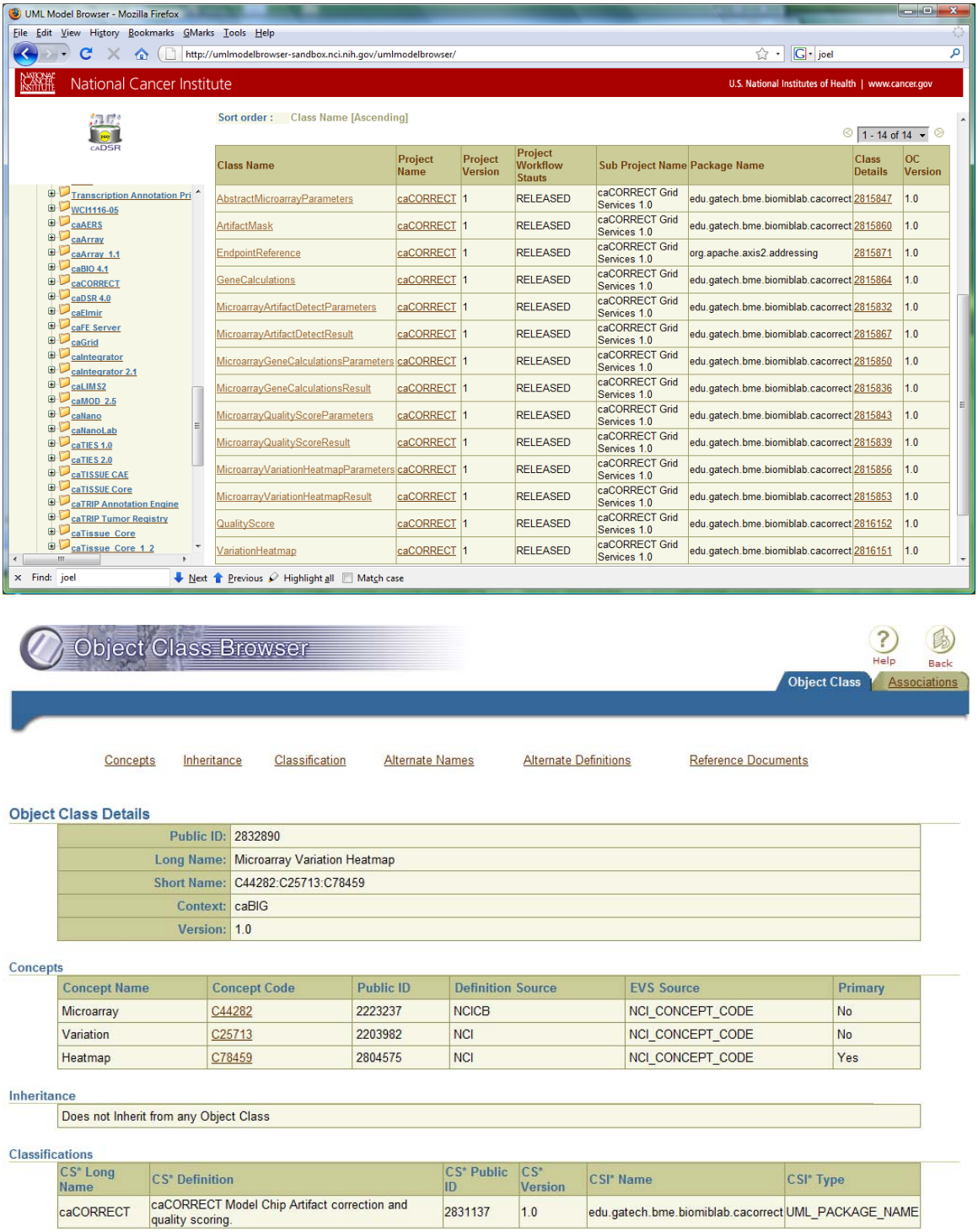


Figure 34: Screenshots of caCORRECT entries in the Cancer Data Standards Repository (caDSR). 14 classes were created in our model and have been officially registered and are now available for re-use by other bioinformatics application developers.

Model-Driven Development and caDSR

Usage of the Unified Modeling Language (UML) to construct a software design at the beginning of software creation allows greater design flexibility and system visibility. The UML model for caCORRECT is shown in Figure 35. In the case of data providing services, the UML model may be used directly to generate Java code using the cancer Common Ontological Representation Environment (caCORE) Software Development Kit (SDK) [170, 182]. The semantic annotation of designed elements relates concepts with CDEs. A single term registered with the Enterprise Vocabulary Service (EVS) is made the primary concept of the CDE (see Table 8 for examples). Additional concepts are added for enhanced clarity and specificity of meaning. A caDSR tool named the Semantic Integration Workbench (SIW) is the most prominent tool to aid in the procedure of annotating a model.

The Enterprise Vocabulary Services (EVS) define a controlled vocabulary, or ontology, for terms specific to the cancer realm [183]. Existing ontologies (e.g. Gene Ontology and Unified Medical Language System (UMLS)) are referenced instead of duplicating work. EVS is regularly updated by merging updates to all the associated ontologies. The Cancer Data Standards Repository (caDSR) is the central registry for Common Data Elements (CDEs).

The Cancer Bioinformatics Infrastructure Objects (caBIO) represent the biomedical domain as a hierarchical object model. These classes are freely available and can save bioinformaticians the trouble of implementing Simple Object Access Protocol (SOAP) and XML application programming interfaces for services that transact with CDEs. An even more exciting benefit of incorporating these classes into web services for new and legacy tools is that the end user application has a layer of protection from updates to data standards.

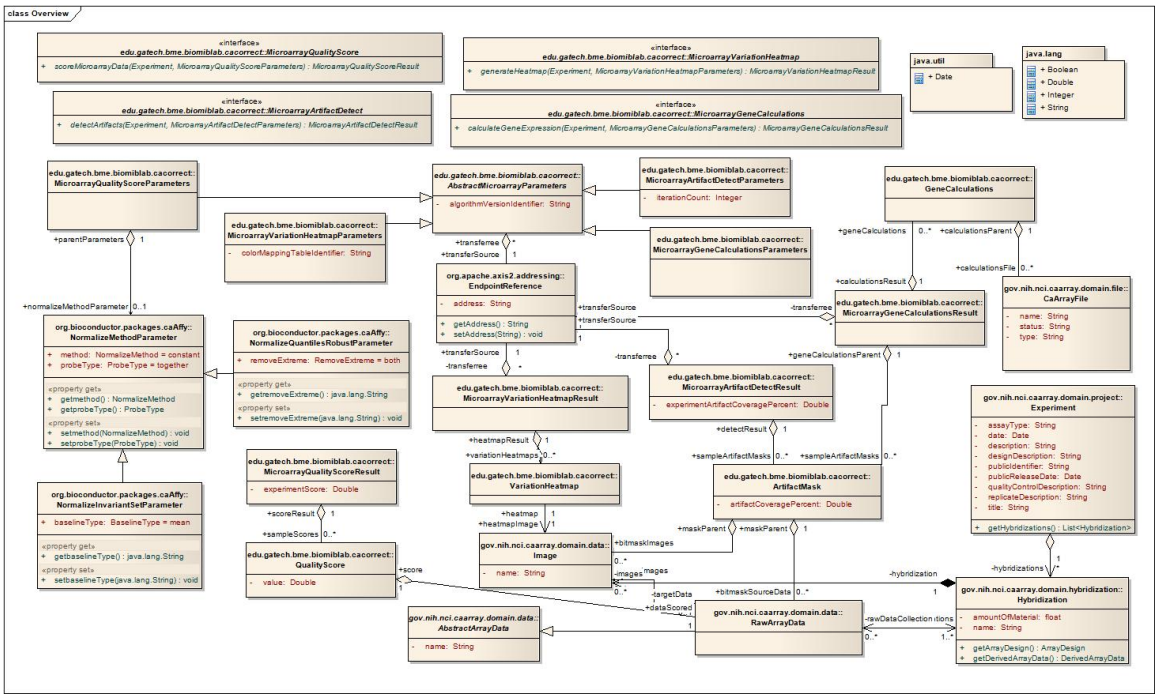


Figure 35: UML Diagram of the caCORRECT analytical grid services. This diagram is a complete description of all input and output data elements for the four caCORRECT grid services: MicroarrayQualityScore, MicroarrayVariationHeatmap, MicroarrayArtifactDetection, and MicroarrayGeneCalculations. All data types have been mapped to standards and all variable and class names have been annotated using a standardized vocabulary. This model reuses 6 data elements from the previously approved caARRAY model and 3 data elements from the Bioconductor model. Four new vocabulary terms were defined in making the model and a new architectural concept, the EndpointReference object was defined. The code and documentation generated for these services matches this model exactly.

Table 8: Concepts re-used or newly defined for the caCORRECT Grid Services.

Term	Definition
Artifact	A structure or appearance that is not naturally present, but has been introduced though manipulation.
Bitmask	**A binary number in which a bit is set to either on or off to store information for bitwise operations.**
Color	The appearance of objects (or light sources) described in terms of a person's perception of their hue and lightness (or brightness) and saturation.
Control	The act of directing or determining; regulation or maintenance of a function or action; a relation of constraint of one entity (thing or person or group) by another.
Cover	Span a region or interval of distance, space or time.
Detection	The activity of perceiving, discerning, discovering or identifying
Heatmap	A graphical representation of data where the values taken by a variable in a two-dimensional map are represented as colors.
Integrity	The state of being complete or undivided, of being sound or undamaged.
Mapping	The creation of a two-dimensional graphic representation of an area or structure, showing the relative position of features or characteristics.
Microarray	A piece of glass or plastic on which different samples have been affixed at separate locations in an ordered manner thus forming a microscopic array. The samples are usually DNA fragments but may also be antibodies, other proteins, or tissues.
Raw	Not processed or refined.
Score	A number or range of numeric values measuring performance, function, quality, or ability.

Community-Reviewed Grid Services for caCORRECT

We have performed numerous studies on the progress of caBIG as an infrastructure project, as well as the progress of various components toward adoption in the caBIG community. One significant effort was the adoption of the tool caNanoLab to begin our involvement. Next we developed a preliminary set of grid services that expose the key elements of the caCORRECT workflow to any developer familiar with the caGrid infrastructure. Four grid service interfaces were developed to correspond to the key functional offerings of caCORRECT: 1) MicroarrayQualityScore, 2) MicroarrayVarianceHeatmap, 3) MicroarrayArtifactDetection, and 4) MicroarrayGeneCalculations (see Figure 36). These grid services are constructed using the Introduce software from caBIG and follows the caBIG compatibility guidelines [184]. These guidelines require well-documented APIs, reuse of existing caBIG infrastructure and data models where possible, and preparation of vocabulary terms for mapping to the Enterprise Vocabulary Services.

Adopting caNanoLab Technology for Laboratory Information Management

A number of preparatory actions were taken in order to smooth the way toward effective involvement with the caBIG initiative. These included attendance at caBIG Annual meetings for three consecutive years, face-to-face meetings with leaders in the Integrative Cancer Research (ICR) workspace (including traveling for a meeting in Boston in October), presentations of an overview of caCORRECT on a scheduled ICR conference call. Most importantly, however, was our demonstration of technical abilities by adopting the caNanoLab application and the demonstration of good will by gathering useful feedback from the members of Dr. Shuming Nie's nanotechnology laboratory for the caNanoLab development team. caNanoLab is a knowledge management application developed by the Nanotechnology Characterization Laboratory and the National Cancer

Institute. This tool is an important repository for source characterization data of nanoparticles [185].

caCORRECT Grid Services System Documentation

Four grid services have been developed, tested, and deployed to the caGRID based on the caCORRECT documentation. These services are currently set up to use default datasets for demonstration purposes. The remaining work on this effort is to integrate our services with the caARRAY Microarray repository so that grid service users may pass any public experiment ID from caARRAY and retrieve caCORRECT analysis results for that experiment. This will require the use of a new caGRID architectural component, called caGridTransfer, and a new API for caARRAY that was released in October 2008.

Throughout the development of preliminary studies, we have followed good software documentation practices. Many of these tools were developed in a team environment, with teams made up of undergraduate students, Master's students, and Ph.D. students sharing the tasks of design, development, and testing. This has forced our tools to operate in a modular fashion, to allow for the use of many different software languages and third-party technologies, and to make use of sensible application programming interfaces (APIs).

The Silver-level Compatibility Review for caBIG required a package containing 12 pieces of documentation about the project under review. These are:

1. Brief description of the data system and its design (a Powerpoint presentation).
2. UML Model (Including a class diagram of data classes and a class diagram with API interfaces).
3. Graphical representation of the UML diagrams.

4. Semantically annotated XMI file (must be able to load into the Semantic Integration Workbench (SIW) tool).

5. Error log from SIW.

6. UML Model Submission Form (basic application profile information).

7. Vocabulary Report.

8. Standards Report.

9. Full CDE Use Report (exported from the cancer Data Standards Repository (caDSR) Common Data Elements (CDE) Browser).

10. An API description document (generally automated JavaDocs are used).

11. Test script(s) demonstrating the use of as many of the API methods as possible.

12. Test log(s) produced from running test scripts.

The generation of Javadocs was the one notable task for generation of a submission package for grid service interoperability certification. It is evident that using the Eclipse IDE's GUI for Javadoc generation was the easiest and most efficient method for creating Javadocs. All class objects designed in Enterprise Architect have a “notes” section of their properties which is translated into Javadoc tags in the Introduce-generated code files. The result is of the typical format as shown in Figure 37. The documentation is linked with hypertext viewable within a web browser.

For the submission package, the last important artifacts relating successful design and implementation of a grid service are the test scripts and test logs. These files contain sample client code required to access your grid service programmatically. Mentors use the submitted package to judge CDEs, vocabulary usage, and architecture. Upon certification, the project's architecture and annotations will be uploaded into the caDSR for use by other projects desiring interoperability. This package was submitted to our caBIG mentors on March 15, 2009 and the review should be complete by April 15, 2009.

Table 9: Summary of community tools under review or already approved for caBIG Silver Compatibility.

Project Name	**Bioinformatics Research Area**	Number of Common Data Elements
BioConductor	Microarray analysis	75
caBIO	Gene-centric data system	369
caIntegrator/Rembrandt	Data source integration	326
caMOD v2.1	Public cancer models	301
caTIES	Tissue pathology reports	219
caTissue Core	Tissue banking, management	326
caTRIP	Grid service workflow	96
caXchange	Clinical trial workflow	19
Function Express	Automated gene annotation	87
GenePattern	Multidisciplinary analysis	187
PIR	Protein data discovery	197
ProteomicsLIMS	Proteomic lab management	208
Reactome	Curated pathway database	81
RProteomics	Mass spec. analysis	40
SEED	Distributed genomic annotation	17
TrAPPS	Mutation experiment management	92
AIM	Image annotation and markup	162
geWorkbench	Integration and visualization	92
IVI middleware	In Vivo Imaging SDK	209
NCIA	Cancer data repository, images	107
PSC	Clinical trial management	67

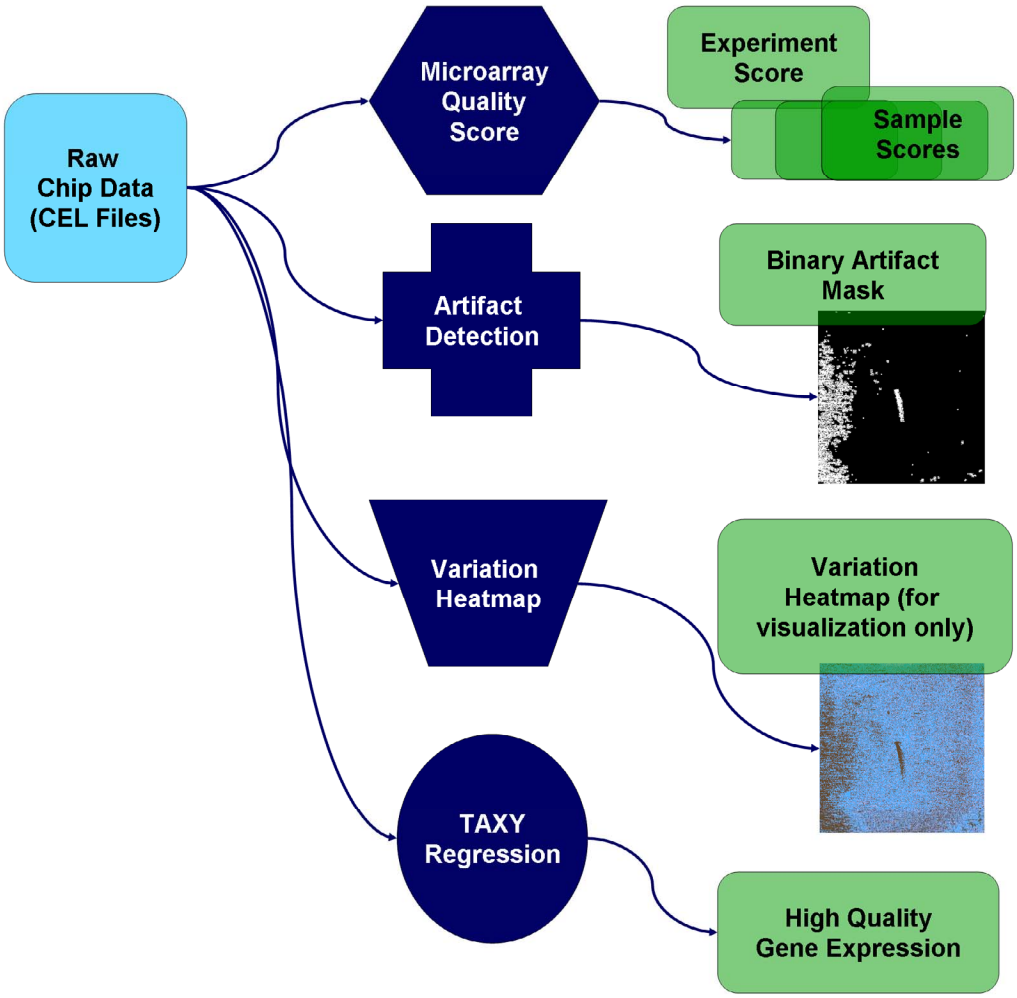


Figure 36: The four grid services developed for caBIG. These functional services were split from the integrated workflow used to drive the caCORRECT web site so that caBIG developers could use them within integrated workflows. The TAXY regression algorithm was developed by Richard Moffitt and Weiguang Wang.

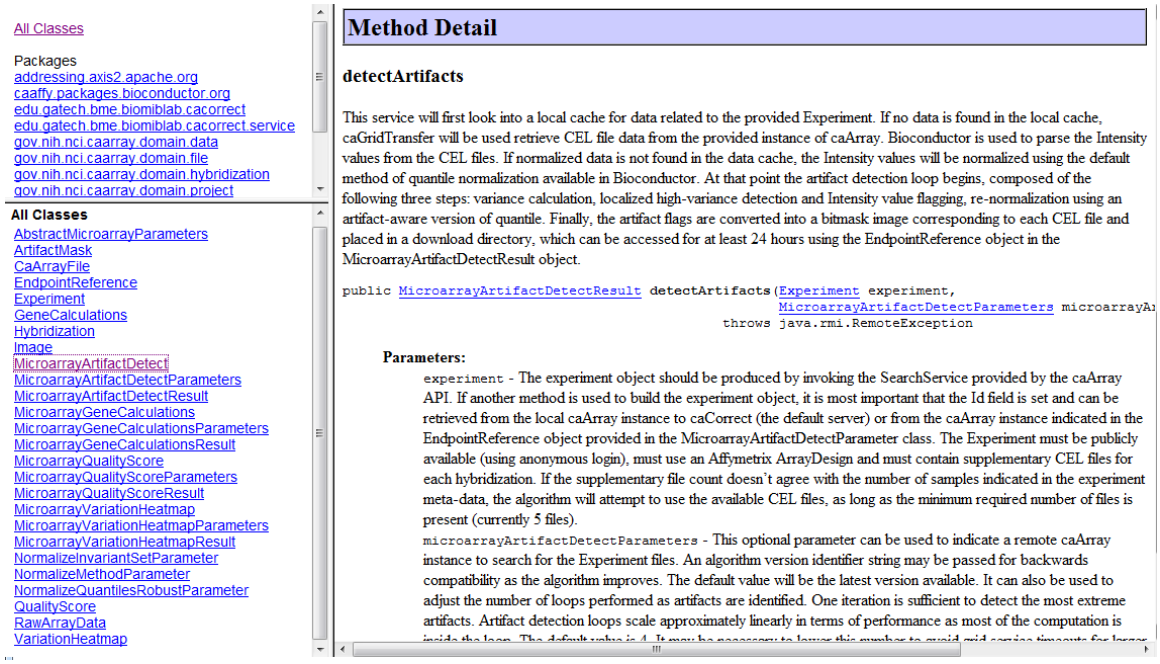


Figure 37: Example documentation (JavaDocs) generated for caBIG compatibility review. The Application Programming Interface (API) documentation describes each of the four analytical services available, as well as what values are expected from the input objects and how to interpret the values provided in the output objects.

CHAPTER 7

CONCLUSION

This dissertation presented an information management and visualization platform for Translational Biomedical Informatics (TBMI). TBMI is a critical step toward personalized medicine because it addresses the inevitable obstacles of information overload and more complex interpretations of data relationships. TBMI also includes development of common infrastructure like that developed by Cancer Biomedical Informatics Grid (caBIG) and standards for data analysis like those developed by the FDA MAQC Phase II Project.

The Concrete Application Deliverables

The concrete goals of this dissertation were to define and demonstrate key technological (or engineering) choices that support the three objectives. These concrete engineering choices are:

1. Development of all software and algorithms using technologies that are easily deployed to the web environment and subsequently a community grid environment such as caBIG.
2. Development of metrics for data quality that can be stored and used through an integrated system for comparison.
3. Use and extension of two technologies that allow for a union of data visualization and the source data used to generate the visualization (assisting with interactive exploration): BioPNG and SVG.

Four application deliverables demonstrate the concrete choices:

1. caCORRECT [18] demonstrates a web-based interactive interface using SVG, AJAX, and JavaScript. It also includes a data quality metric.

2. ArrayWiki [68] is a web-based open editing system for community contribution. ArrayWiki exposes the data quality metrics in caCORRECT in a searchable repository. It natively supports storage of SVG and PNG visualizations in a flexible framework.

3. BioPNG [68] unites source data with visualization by intelligently fragmenting large data values for storage into a 8 or 16-bin 4-channel compressed format. A web-based interface allows users to create their own files from source data.

4. SimpleVisGrid [186] unites source data with visualization by converting certain biological source data into SVG files or BioPNG. This application is grid-ready and properly modeled for integration with caBIG.

In addition to application deliverables, two other concrete deliverables were accomplished to support the objective of translation. There were:

1. Participation in the FDA MAQC Phase II Project (as evidenced by authorship in the special issue publication).
2. Submission of a Silver-level Review package for certification of caCORRECT Grid Services.

caCORRECT

The chip artifact Correction Tool (caCORRECT, http://cacorrect.bme.gatech.edu) is a web-based application for identifying data problems on many types of microarrays. caCORRECT also includes four caBIG-approved Grid Services for use by the wider cancer community in larger experimental workflows. caCORRECT supports Affymetrix GeneChip and SNP microarrays. The three major features of caCORRECT are the quality score for comparing data sets, the visualization heatmap of chip variance for uncovering lab protocol problems, and the calculation of artifact-aware gene expression results for more reproducible biomarker results.

ArrayWiki

ArrayWiki (http://arraywiki.bme.gatech.edu) is a community-maintained microarray repository. ArrayWiki is different from other microarray repositories primarily in the fact that any registered user of the system may edit, add, or re-organize experiments within the repository. Other unique features are the inclusion of all quality results produced by caCORRECT and a unique data storage format, BioPNG, that combines compression and visualization and offers improved security over zip files used by other repositories. ArrayWiki contains over 10,000 microarray chips and is steadily growing on a daily basis.

BioPNG and Scalable Vector Graphics

BioPNG is an open-source data storage algorithm that combines very high lossless compression rates with the ability to view data files in nearly all standard image viewers or editors in addition to web browsers. For this reason, it is highly portable and allows users to easily verify common corruption issues that may be missed when formats that do not offer visualization are used. BioPNG is flexible enough to support very large integers in addition to very precise decimal values as well as three-dimensional data using the Animated PNG (APNG) format.

SimpleVisGrid is a collection of grid-enabled visualization services that all offer the unique feature of wrapping source data into the same file format with the visualization data. This allows service consumers (e.g. viewer applications) to manipulate the underlying data and thus the resulting visualization. This creates more useful and engaging graphics because users can explore the data. It also improves the quality of communications using data because users can verify assumptions and modify graphics to better suit their own particular context.

Food and Drug Administration Microarray Quality Control Consortium

The Microarray Quality Control (MAQC) Consortium of the Food and Drug Administration (FDA) Center for Toxicological Research and participating institutions is a four phase project. Phase I was published in *Nature Biotechnology* in 2006. The purpose of the Phase I effort was to determine if microarrays were technically reliable as a data acquisition technology. The Consortium published a set of standards for researchers to follow when performing microarray experiments to support New Drug Applications (NDAs). The purpose of Phase II is to determine that data analysis protocols for mining microarray data are reliable when applied to different clinical datasets. This effort will determine if microarray technology can itself be approved for diagnostic or prognostic purposes. The secondary goal is to publish standards for building the necessary clinical classification models. This dissertation will discuss an effort to exhaustively compare all of the properties of the K Nearest Neighbors (KNN) classifier as a contribution toward this goal. After Phase II is complete, the MAQC Consortium will apply the same rigorous study to Next Generation Sequencing (NGS) technologies such as Roche 454, Illumina Genome Analyzer, Applied Biosystems Supported Oligo Ligation Detection (SOLiD) technology and Helicos Biosciences HeliScope Sequencer.

Cancer Biomedical Informatics Grid (caBIG) Certified Services

Four grid service interfaces were developed to correspond to the key functional offerings of caCORRECT: 1) MicroarrayQualityScore, 2) MicroarrayVarianceHeatmap, 3) MicroarrayArtifactDetection, and 4) MicroarrayGeneCalculations. A fully-annotated UML model accompanied these services and all of the input and output data elements are registered in the Cancer Data Standards Repository (caDSR). A submission package was prepared for the Silver-level certification review committee. This package included JavaDoc documentation, a testing plan and test scripts in addition to the UML model. A decision is expected on the certification in early April.

Future Impacts of this Work

Of the key contributions of this work, some may seem to be obvious extensions of well-known "good science" principles. For example, the importance of having a very deep understanding of the data that you are working with has been stressed since the early days of data analysis. However, it is important to note that the preponderance of data available today combined with the increasingly complex methods used for high-throughput data acquisition mean that scientists must become increasingly specialized in order to accomplish this deep understanding. This specialization of many scientists creates a need for better systems and more standardization to help the specialists work together on team projects. The problems of modern science are no longer solvable by an independent researcher. Participation in "Team Science" and embracing this new culture of scientific research has been one of the most inspiring and gratifying parts of working on this dissertation.

Many well-known Internet gurus such as Tim Berners-Lee are saying that the next 10 years will be marked by an enormous rise in "linked data" as contrasted to the previous emergence of "linked documents." Tim O'Reilly calls this new phase of Internet evolution "Web 2.0." I am proud to say that ArrayWiki displays all of the characteristics of a Web 2.0 Platform as spelled out by the authors of Wikipedia (see http://en.wikipedia.org/wiki/Web_2.0). I think that use of image formats to pass raw data in open transactions will become another important feature of Web 2.0 that does not yet have full recognition. I never cease to be amazed by the inventive minds of my fellow software developers and of the rapid pace of innovation spurred on by increased connectivity of such able minds. I hope that this work can inspire some of those young minds to direct their passion at solving some of the greatest problems yet: human disease and how to deliver effective health care to treat it.

APPENDIX A

RELEVANT PUBLICATIONS COMPOSING THIS DISSERTATION

In Preparation/Submitted

Stokes TH, Hang S, Wang MD, "SimpleVisGrid: Grid-based Visualization Services for Translational Biomedical Informatics," (In preparation).

The MAQC Consortium (currently 159 authors led by Leming Shi). "The MAQC-II Project: A comprehensive study of common practices for the development and validation of microarray-based predictive models," *Nature Biotechnology*. (Under review for MAQC Phase II Special Issue, June).

Parry RM, Jones W, Stokes TH, Phan JH, Moffitt RA, Fang H, Shi L, Oberthuer A, Fischer M, Tong W, Wang MD, "K-nearest neighbors (KNN) models for microarray gene-expression analysis and reliable clinical outcome prediction", *Nature Biotechnology*. (Under review for MAQC Phase II Special Issue, June).

Jones W, Wang MD, Moffitt RA, Phan JH, Stokes TH, Bao W, Wolfinger R, Li L, Parker J. "The impact of quality-related artifacts on the predictive performance of Affymetrix GeneChips in a diagnostic setting: a case study within MAQC-II." *Nature Biotechnology*. (Under review for MAQC Phase II Special Issue, June).

Osunkoya AO, Yin-Goen Q, Phan JH, Moffitt RA, Stokes TH, Wang MD, Young AN. "Diagnostic biomarkers for renal cell carcinoma: identification with novel bioinformatics systems for microarray data analysis." *Human Pathology*. (Under review).

Journal/Book Publications

Stokes TH, Moffitt RA, Phan JH, and Wang MD, “chip artifact CORRECTion (ca-CORRECT): a bioinformatics system for quality assurance of genomics and proteomics array data," Annals of Biomedical Engineering, vol. 35, pp. 1068-1080, 2007.

Stokes T., Torrance, J.T., Li, H., Wang, M.D. ArrayWiki: an enabling technology for sharing public microarray data repositories and meta-analyses. BMC Bioinformatics, 2008 Jun; 9(Suppl 6):S18.

Merrill AH Jr, Stokes TH, Momin A, Park H, Portz BJ, Kelly S, Wang E, Sullards MC, Wang MD. “Sphingolipidomics: a valuable tool for understanding the roles of sphingolipids in biology and disease.” Journal of Lipids Research. Nov. 21 2008.

Phan JH, Moffitt RA, Stokes TH, Liu J, Young AN, Nie S, Wang MD. “Biomarkers, Nanotechnology, and Personalized Medicine.” *Trends in Biotechnology*. 2009.

Conference Proceedings

T. H. Stokes, J. T. Torrance, N. L. Goasduff, H. Li, and M. D. Wang, “Arraywiki: Liberating Microarray Data from Non-Collaborative Public Repositories.” *International Multi-Symposiums on Computer and Computational Sciences, IMSCCS*, Iowa City, IA, 2007, pp. 92-99.

J. H. Phan, R. A. Moffitt, T. H. Stokes, and M. D. Wang, “Evolving Biological Behavior in Gene-Based Cellular Simulations," in *IEEE 7th International Symposium on BioInformatics and BioEngineering, BIBE*, Boston, MA, 2007. pp. 509-516.

T. H. Stokes, R. X. Han, R. A. Moffitt, and M. D. Wang, “Extending Microarray Quality Control and Analysis Algorithms to Illumina Chip Platform.” *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, France, 2007, pp. 4637-4640.

J. T. Torrance, R. A. Moffitt, T. H. Stokes, and M. D. Wang, “Can We Trust Biomarkers? Visualization and Quantification of Outlier Probes in High Density Oligonucleotide Microarrays.” *IEEE/NIH Life Science Systems and Applications Workshop, LSSA*, Boston, MA, 2007, pp. 196-199.

T. H. Stokes, J. H. Phan, C. F. Quo, S. Nie, and M. D. Wang, “Bio-Nano-Informatics: An Integrated Information Management System for Personalized Oncology," in *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, New York City, NY, 2006, pp. 3325-3328.

T. H. Stokes, J. H. Phan, W. M. Feng, G. Tuteja, and M. D. Wang, “GAVis: a Tool for Visualization and Control of Genetic Algorithms for -omic Data Analysis," in *27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Shanghai, China, 2005, pp. 2855-2858.

P. Henning, D. Stiles, T. H. Stokes, G. Wang, D. Wheeler, I. Sidorov, P. Tan, M. Cam and M. D. Wang, “ChipQC: Microarray Artifact Visualization Tool.” *Int'l Conf. on Intelligent Systems for Molecular Biology*, 2005 Jun 25-29.

R. A. Moffitt, T. H. Stokes, J. H. Phan and M. D. Wang, “Simple Outlier Removal Improves the Results of Support Vector Machines as a Biomarker Selection Tool.” *Int'l Conf. on Intelligent Systems for Molecular Biology*, 2005 Jun 25-29.

APPENDIX B

GLOSSARY OF TERMS

Adaptability – a measure of the projected useful lifetime of a computational solution. This factor is greatly influenced by behind-the-scenes design decisions, such as the development, database, and deployment platforms and also the modularity of the code. Each enhancement requested by users will cause the development team to consider a complete rewrite of the system. Meanwhile, the development team will experience turnover, with new members bringing on new skills. Over time, the likelihood of a complete rewrite approaches inevitability. However, useful and adaptable solutions that require fairly low maintenance tend to transition over time from one solution to the next.

AJAX – Asynchronous JavaScript and XML, a technique for making web pages more interactive by allowing small modules within the page to exchange information with the server without rendering the entire page each time. AJAX is not a standalone technology, but rather a grouping of existing technologies used in a special way.

Batch effect - The effect of "the day of the assay" on microarray gene expression. In other words, microarray data from a particular day, or batch of reagent, may have a range of expression values that differs from other days or reagents. Without checking for a batch effect, a microarray dataset may suggest a set of differentially expressed genes reflects clinical categories, when in fact the genes are all artifacts. It is important to plan for batch effects when planning how samples will be analyzed. Samples from different clinical categories should not be analyzed together on separate days. Rather, it is better to run samples from different clinical categories in the same batch.

BioPAX - Biological Pathway data-exchange. An RDF-based biological pathway data exchange format. The current release is Level 2, version 1.0. This version integrates Reactome (http://reactome.org/). BioPAX began at the Fourth BioPathways Consortium Meeting at Intelligent Systems in Molecular Biology (ISMB) August 2002 in Edmonton, Canada.

BPEL4WS – Business Process Execution Language for Web Services, a language for formal specification of business interactions. BPEL4WS is intended to expand the role of automated process integration in corporate and business-to-business space. It is an example of workflow creation technology that may be extended to biomedical research workflows.

caBIG – Cancer Bioinformatics Grid, the proposed World Wide Web of cancer research, steered and developed primarily at the National Cancer Institute (NCI) of the National Institutes of Health (NIH) branch of the United States government.

caBIO - Cancer Bioinformatics Infrastructure Objects, an Application Programming Interface (API) developed first as Java objects but extended to support various programming platforms. These objects represent the most fundamental concepts in bioinformatics research, such as Gene, Ontology, and Sequence.

caDSR - Cancer Data Standards Repository, a Data Dictionary for bioinformatics research. The purpose of caDSR is for data definitions in bioinformatics software to become uniform. For example, if a Gene data object is defined in 34 approved caBIG applications, that data object should contain uniform fields, properties, and methods across these applications to aid in understanding and interoperability. If your application uses caBIO objects exclusively, you automatically fulfill this requirement.

CDE - Common Data Element, a term for each data definition in caDSR.

Cross-validation - A method of repeatedly partitioning the data into separate training and testing sets. Because of these reiterations, every sample becomes used in training and testing capacities. To avoid introducing bias, the biostatistician can resort to use of a holdout set.

DTD - Document Type Definition, a fundamental XML concept that preceded the XSD. DTDs simply indicated the expected structure of an incoming XML document and could be used for validation.

EVS - Enterprise Vocabulary Services, a caBIG ontology unification initiative intended to provide a one-stop resource for cancer-related word definitions. EVS incorporates many other medical ontologies.

External validation - Validation of a microarray predictor using a dataset that was not used in the development of the microarray predictor.

GO - Gene Ontology, an early effort in biological ontologies. The goal of GO is to capture relationships between discoveries about the functional roles of genes in organisms in an hierarchical structure.

HTML - HyperText Markup Language, the basic communication language of the World Wide Web. HTML is intended to format data for visual appeal to human users, as opposed to XML, which is data formatting for software consumption.

Interoperability – a measure of how well a system developed by one team of people can coordinate workflows with all other systems in use by the user community. Interoperability is too often applied only to tools when what is really the currency of the underlying interoperability of tools is the data. So while data standards are incorporated at the tool level, the development of those standards is independent of tools and only considers data requirements of the domain.

LSID - Life Science Identifier (http://sourceforge.net/projects/lsid/), under the oversight of the Interoperable Informatics Infrastructure Consortium (I3C) and IBM, LSID is intended to provide a uniform way to name and locate life science data. The vision of the LSID is that computational scientists could associate them with applications for visualization and analysis, thus improving semantic integration.

MAGE-ML – Microarray Gene Expression Markup Language, an XML schema resulting from the call for more meta-data about microarray experiments to improve comparison and validate reproducibility. MAGE-ML is now integrated into many well-established microarray databases.

MathML – Mathematics Markup Language, an XML schema for visual representation of mathematical equations. MathML can be parsed by software which interprets and uses the equations for analysis and modeling, but does not specify how an equation should be used in its specification.

Matthew's correlation coefficient - A performance measure of microarray predictiveness. The Pearson correlation coefficient provides the same number for binary data.

MyGrid – (not an acronym), a bioinformatics workflow integration effort originating in the UK. MyGrid uses the Scufl workflow description language and the Taverna workbench user interface for building the workflows. They current hot project is the myExperiment portal for sharing workflows among the community.

Normalization - The process of aligning gene expression measures from different microarrays, so that a comparison can occur. The goal of normalization is to avoid spurious differences between sample categories. Normalization traditionally consists of two components: 1) mean centering (in which the global mean expression value of all the probes in the microarrays are calculated and the differences in these means subtracted) 2) scaling (in which gene expression values after mean centering are divided by the estimated global variance). Quantile normalization, in which normalization proceeds through ranked gene expression quantiles of the microarray data, is an example. Normalization, itself, can introduce spurious findings. Therefore, some microarray analysis plans may incorporate an examination of the data without normalization.

OWL - Web Ontology Language, an XML schema for defining ontologies. A great result of OWL is to aid the integration of ontologies from different sources. OWL incorporates concepts defined by RDF.

Pre-analysis - A "first look" at the data to identify unusual situations and as a quality control check. Pre-analysis tasks may include outlier identification, batch effect determination, and determination if a statistically significant difference in gene expression exists between clinical categories (prior to modeling).

RDF - Resource Description Framework, an XML schema for defining tertiary relationships (i.e. subject, predicate, object). Each of the three elements of the tertiary relationship is considered a resource and thus a hub of other tertiary relationships. RDF is now in widespread use as a fundamental semantic structure for more specific efforts toward software that "understands" data.

Reliability - The similarity between performances of a predictive model as estimated during cross-validation and performance of that predictive model during external validation. Or, the similarity between performance of a DAP as estimated on one dataset, and performance of that DAP on a new dataset (including swapping of training and test data).

Reproducibility - The similarity between performance of a DAP as conducted by the original DAP authors and performance of a reimplementation of that DAP by a different team.

Semantic Web – A proposed extension to the existing World Wide Web where web-based documents and content contain machine-readable instructions for how computers should interpret the meaning of the content. The goal of the Semantic Web is to increase the usability of the Internet by allowing a user to execute automated, highly complex goal-based searches, thus reducing the amount of time spent "surfing".

SBML – Systems Biology Markup Language, an XML schema for defining biochemical reaction models in systems biology. SBML is in widespread use and is supported by more than 70 systems biology modeling software packages.

SBGN – Systems Biology Graphics Notation, a collection of symbols for the presentation of biochemical reaction networks. The symbolic language is not intended to convey a mathematically precise description of the behavior of the network, but rather to give a qualitative description.

SVG - Scalable Vector Graphics, an XML schema for two dimensional graphics representation. SVG defines basic elements such as shapes, images, gradients and symbols (complex shapes). SVG is natively supported by most modern web browsers, allowing developers to create richly-varied user interfaces, rather than the standard HTML forms offered up to this point by all web sites. SVG is most easily compared to Java applets or Macromedia Flash from the user perspective, but is actually vastly different in the sense that it is open and scalable.

UDDI - Universal Description, Discovery, and Integration directory, a directory of web services. UDDI is a cross-industry effort driven by major platform and service providers within the OASIS standards consortium. UDDI can be used in a single-registry or distributed-registry approach, but does not specify how interactions between registries might take place.

UMLS - Unified Medical Language System, an ontology of medical terms with cross-references. UMLS is one of the source ontologies of caBIG's EVS.

URN - Universal Resource Names, the Web addressing scheme that allows a text string to be used to identify and locate resources on the network. URN is synonymous with the more common URLs published by any groups with a web presence.

WSDL - Web Services Description Language, an XML schema for identifying the purpose and interaction schemes of web services. WSDL does not describe web service location strategies but is a successful standard in that all location strategies have adopted WSDL for description.

XML - Extensible Markup Language, the fundamental technology behind most text-based data identification and location schemes. Some might say that HTML is just another implementation of an XML schema, although it came into widespread use much earlier.

XSD - XML Schema Definition, a text document that specifies the structure of other XML documents. XSDs do not contain any data, only meta-data such as names and types and hierarchical structure. XSDs are the successor to DTDs and provide more precise definition for the purposes of validation.

Usability – a measure of usefulness of a computational solution. This is often measured directly by user satisfaction surveys, but may also be aggregated with quantitative measures of complexity (number of lines of code, screens, buttons, menus) and the raw growth of the user base (e.g. the users may be largely dissatisfied, but the demand is so great because the functionality is needed widely and there is no substitute).

REFERENCES

[1] J. Kaiser, "Cancer research - Von Eschenbach revises the NCI agenda," *Science,* vol. 303, pp. 1952-1952, Mar 26 2004.

[2] H. A. Piwowar and W. Chapman, "Identifying data sharing in biomedical literature," *AMIA Annu Symp Proc,* pp. 596-600, 2008.

[3] H. A. Piwowar, M. J. Becich, H. Bilofsky, and R. S. Crowley, "Towards a Data Sharing Culture: Recommendations for Leadership from Academic Health Centers," *PLoS Med,* vol. 5, p. e183, Sep 2 2008.

[4] M. Gardner, "The fantastic combinations of John Conway's new solitaire game "life"," in *Scientific American*, 1970, pp. 120-123.

[5] E. Birney, A. Bateman, M. E. Clamp, and T. J. Hubbard, "Mining the draft human genome," *Nature,* vol. 409, pp. 827-8, Feb 15 2001.

[6] B. J. Barnhart, "DOE Human Genome Program," in *Human Genome Quarterly*. vol. Spring, 1989.

[7] "The human genome. Science genome map," *Science,* vol. 291, p. 1218, Feb 16 2001.

[8] M. Kanehisa and P. Bork, "Bioinformatics in the post-sequence era," *Nat Genet,* vol. 33 Suppl, pp. 305-10, Mar 2003.

[9] I. Foster, "Service-oriented science," *Science,* vol. 308, pp. 814-817, May 6 2005.

[10] L. Stein, "Creating a bioinformatics nation - A web-services model will allow biological data to be fully exploited.," *Nature,* vol. 417, pp. 119-120, May 9 2002.

[11] "Need to Analyse Voluminous Data Opens up Exciting Opportunities for Bioinformatics," Frost & Sullivan November 21 2005 2005.

[12] G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki, "DAVID: Database for annotation, visualization, and integrated discovery," *Genome Biology,* vol. 4, pp. -, 2003.

[13] B. Staats, L. Qi, M. Beerman, H. Sicotte, L. A. Burdett, B. Packer, S. J. Chanock, and M. Yeager, "Genewindow: an interactive tool for visualization of genomic variation," *Nat Genet,* vol. 37, pp. 109-10, Feb 2005.

[14] T. Toyoda, Y. Mochizuki, K. Player, N. Heida, and Y. Sakaki, "OmicBrowse: a browser of multidimensional omics annotations," *Bioinformatics,* Oct 31 2006.

[15] T. Toyoda and A. Wada, "Omic space: coordinate-based integration and analysis of genomic phenomic interactions," *Bioinformatics,* vol. 20, pp. 1759-1765, Jul 22 2004.

[16] D. B. Finkelstein, "Trends in the Quality of Data from 5168 Oligonucleotide Microarrays from a Single Facility," *Journal of Biomolecular Techniques: JBT,* vol. 16, p. 143, 2005.

[17] [Anon], "The database revolution," *Nature,* vol. 445, pp. 229-230, Jan 18 2007.

[18] T. H. Stokes, R. A. Moffitt, J. H. Phan, and M. D. Wang, "chip artifact CORRECTion (caCORRECT): A Bioinformatics System for Quality Assurance of Genomics and Proteomics Array Data," *Ann Biomed Eng,* vol. 35, pp. 1068-80, Jun 2007.

[19] O. Ritter, P. Kocab, M. Senger, D. Wolf, and S. Suhai, "Prototype implementation of the integrated genomic database," *Comput Biomed Res,* vol. 27, pp. 97-115, Apr 1994.

[20] S. L. Salzberg, "Genome re-annotation: a wiki solution?," *Genome Biol,* vol. 8, p. 102, 2007.

[21] L. D. Stein, "Integrating biological databases," *Nature Reviews Genetics,* vol. 4, pp. 337-345, May 2003.

[22] K. H. Cheung, K. Y. Yip, A. Smith, R. deKnikker, A. Masiar, and M. Gerstein, "YeastHub: a semantic web use case for integrating data in the life sciences domain," *Bioinformatics,* vol. 21, pp. I85-I96, Jun 2005.

[23] T. Liefeld, M. Reich, J. Gould, P. L. Zhang, P. Tamayo, and J. P. Mesirov, "GeneCruiser: a web service for the annotation of microarray data," *Bioinformatics,* vol. 21, pp. 3681-3682, Sep 15 2005.

[24] W. Tong, X. Cao, S. Harris, H. Sun, H. Fang, J. Fuscoe, A. Harris, H. Hong, Q. Xie, R. Perkins, L. Shi, and D. Casciano, "ArrayTrack--supporting toxicogenomic research at the U.S. Food and Drug Administration National Center for Toxicological Research," *Environ Health Perspect,* vol. 111, pp. 1819-26, Nov 2003.

[25] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang, "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biol,* vol. 5, p. R80, 2004.

[26] A. Birkland and G. Yona, "BIOZON: a system for unification, management and analysis of heterogeneous biological data," *BMC Bioinformatics,* vol. 7, pp. -, Feb 15 2006.

[27] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and Isdn Systems,* vol. 30, pp. 107-117, Apr 1998.

[28] T. Clark, S. Martin, and T. Liefeld, "Globally distributed object identification for biological knowledgebases," *Brief Bioinform,* vol. 5, pp. 59-70, Mar 2004.

[29] J. P. Ioannidis, D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, J. Mangion, T. Mehta, M. Nitzberg, G. P. Page, E. Petretto, and V. van Noort, "Repeatability of published microarray gene expression analyses," *Nat Genet,* vol. 41, pp. 149-55, Feb 2009.

[30] A. Dupuy and R. M. Simon, "Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting," *JNCI Journal of the National Cancer Institute,* vol. 99, p. 147, 2007.

[31] R. Simon, M. Radmacher, K. Dobbin, and L. McShane, "Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification," *Journal of the National Cancer Institute,* vol. 95, pp. 14-18, 2003.

[32] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, and H. C. Causton, "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data," *NATURE GENETICS,* vol. 29, pp. 365-372, 2001.

[33] N. Le Novere, A. Finney, M. Hucka, U. S. Bhalla, F. Campagne, J. Collado-Vides, E. J. Crampin, M. Halstead, E. Klipp, P. Mendes, P. Nielsen, H. Sauro, B. Shapiro, J. L. Snoep, H. D. Spence, and B. L. Wanner, "Minimum information requested in the annotation of biochemical models (MIRIAM)," *Nat Biotechnol,* vol. 23, pp. 1509-15, Dec 2005.

[34] L. Shi, L. H. Reid, W. D. Jones, and R. Shippy, "MAQC Consortium. The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements," *Nat Biotechnol,* vol. 24, pp. 1151-1161, 2006.

[35] J. Brettschneider, F. Collin, B. M. Bolstad, and T. P. Speed, "Quality Assessment for Short Oligonucleotide Microarray Data. Rejoinder," *Technometrics,* vol. 50, p. 279, 2008.

[36] R. A. Irizarry, D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. N. Garcia, J. Geoghegan, and G. Germino, "Multiple-laboratory comparison of microarray platforms," *Nature Methods,* vol. 2, pp. 345-350, 2005.

[37] M. L. T. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar, "Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations." vol. 97: National Acad Sciences, 2000, pp. 9834-9839.

[38] K. Garwood, T. McLaughlin, C. Garwood, S. Joens, N. Morrison, C. F. Taylor, K. Carroll, C. Evans, A. D. Whetton, S. Hart, D. Stead, Z. Yin, A. J. Brown, A. Hesketh, K. Chater, L. Hansson, M. Mewissen, P. Ghazal, J. Howard, K. S. Lilley, S. J. Gaskell, A. Brass, S. J. Hubbard, S. G. Oliver, and N. W. Paton, "PEDRo: a database for storing, searching and disseminating experimental proteomics data," *BMC Genomics,* vol. 5, p. 68, Sep 17 2004.

[39] E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer," *Lancet,* vol. 359, pp. 572-7, Feb 16 2002.

[40] J. M. Sorace and M. Zhan, "A data review and re-assessment of ovarian cancer serum proteomic profiling," *BMC Bioinformatics,* vol. 4, p. 24, Jun 9 2003.

[41] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science,* vol. 286, pp. 531-7, Oct 15 1999.

[42] L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. A. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K. Lobenhofer, R. K. Puri, U. Schrf, J. Thierry-Mieg, C. Wang, M. Wilson, P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X. M. Cao, T. A. Cebula, J. J. Chen, J. Cheng, T. M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X. H. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle, S. C. Harris, E. Hatchwell, C. A. Hauser, S. Hester, H. Hong, P. Hurban, S. A. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. LeClerc, S. Levy, Q. Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. A. Patterson, R. G. Perkins, E. H. Peters, R. Peterson, K. L. Philips, P. S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B. A. Rosenzweig, R. R. Samaha, M. Schena, G. P. Schroth, S. Shchegrova, D. D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K. L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S. J. Walker, S. J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, L. Zhang, S. Zhong, Y. Zong and W. Slikker, Jr., "The MicroArray Quality Control (MAQC) project shows inter- and

intraplatform reproducibility of gene expression measurements," *Nat Biotechnol,* vol. 24, pp. 1151-61, Sep 2006.

[43] W. P. Kuo, F. Liu, J. Trimarchi, C. Punzo, M. Lombardi, J. Sarang, M. E. Whipple, M. Maysuria, K. Serikawa, and S. Y. Lee, "A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies," *Nature Biotechnology,* vol. 24, pp. 832-840, 2006.

[44] C. Li and W. H. Wong, "Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 98, pp. 31-36, Jan 2 2001.

[45] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, "Summaries of affymetrix GeneChip probe level data," *Nucleic Acids Research,* vol. 31, pp. -, Feb 15 2003.

[46] M. Suarez-Farinas, A. Haider, and K. M. Wittkowski, ""Harshlighting" small blemishes on microarrays," *Bmc Bioinformatics,* vol. 6, Mar 22 2005.

[47] M. Reimers and J. N. Weinstein, "Quality assessment of microarrays: Visualization of spatial artifacts and quantitation of regional biases," *Bmc Bioinformatics,* vol. 6, pp. -, Jul 1 2005.

[48] J. Brettschneider, F. Collin, B. M. Bolstad, and T. P. Speed, "Quality assessment for short oligonucleotide microarray data," *(unpublished, submitted to Technometrics),* 2006.

[49] J. Gollub, C. A. Ball, G. Binkley, J. Demeter, D. B. Finkelstein, J. M. Hebert, T. Hernandez-Boussard, H. Jin, M. Kaloper, J. C. Matese, M. Schroeder, P. O. Brown, D. Botstein, and G. Sherlock, "The Stanford Microarray Database: data access and quality assessment tools," *Nucleic Acids Research,* vol. 31, pp. 94-96, Jan 1 2003.

[50] R. A. Irizarry, D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. N. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Q. Yang, S. Q. Ye, and W. Yu, "Multiple-laboratory comparison of microarray platforms," *Nature Methods,* vol. 2, pp. 345-349, May 2005.

[51] Y. Kluger, H. Yu, J. Qian, and M. Gerstein, "Relationship between gene co-expression and probe localization on microarray slides," *BMC Genomics,* vol. 4, p. 49, Dec 10 2003.

[52] A. J. Holloway, R. K. van Laar, R. W. Tothill, and D. D. L. Bowtell, "Options available - from start to finish - for obtaining data from DNA microarrays," *Nature Genetics,* vol. 32, pp. 481-489, Dec 2002.

[53] G. A. Churchill, "Fundamentals of experimental design for cDNA microarrays," *Nature Genetics,* vol. 32, pp. 490-495, Dec 2002.

[54] L. Brodsky, A. Leontovich, M. Shtutman, and E. Feinstein, "Identification and handling of artifactual gene expression profiles emerging in microarray hybridization experiments," *Nucleic Acids Research,* vol. 32, pp. -, Feb 2004.

[55] T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W. C. Ngau, P. Ledoux, D. Rudnev, A. E. Lash, W. Fujibuchi, and R. Edgar, "NCBI GEO: mining millions of expression profiles - database and tools," *Nucleic Acids Research,* vol. 33, pp. D562-D566, Jan 1 2005.

[56] H. Parkinson, U. Sarkans, M. Shojatalab, N. Abeygunawardena, S. Contrino, R. Coulson, A. Farne, G. G. Lara, E. Holloway, M. Kapushesky, P. Lilja, G. Mukherjee, A. Oezcimen, T. Rayner, P. Rocca-Serra, A. Sharma, S. Sansone, and A. Brazma, "ArrayExpress - a public repository for microarray gene expression data at the EBI," *Nucleic Acids Research,* vol. 33, pp. D553-D555, Jan 1 2005.

[57] K. Hede, "Cancer data coming soon to laptops everywhere," *J Natl Cancer Inst,* vol. 97, pp. 876-8, Jun 15 2005.

[58] Y. Tateno and K. Ikeo, "[International public gene expression database (CIBEX) and data submission]," *Tanpakushitsu Kakusan Koso,* vol. 49, pp. 2678-83, Dec 2004.

[59] K. Ikeo, J. Ishi-i, T. Tamura, T. Gojobori, and Y. Tateno, "CIBEX: center for information biology gene expression database," *C R Biol,* vol. 326, pp. 1079-82, Oct-Nov 2003.

[60] C. A. Ball, I. A. B. Awad, J. Demeter, J. Gollub, J. M. Hebert, T. Hernandez-Boussard, H. Jin, J. C. Matese, M. Nitzberg, F. Wymore, Z. K. Zachariah, P. O. Brown, and G. Sherlock, "The Stanford Microarray Database accommodates additional microarray platforms and data formats," *Nucleic Acids Research,* vol. 33, pp. D580-D582, Jan 1 2005.

[61] T. L. Fare, E. M. Coffey, H. Y. Dai, Y. D. D. He, D. A. Kessler, K. A. Kilian, J. E. Koch, E. LeProust, M. J. Marton, M. R. Meyer, R. B. Stoughton, G. Y. Tokiwa, and Y. Q. Wang, "Effects of atmospheric ozone on microarray data quality," *Analytical Chemistry,* vol. 75, pp. 4672-4675, Sep 1 2003.

[62] S. Horan, I. Bourges, and B. Meunier, "Transcriptional response to nitrosative stress in Saccharomyces cerevisiae," *Yeast,* vol. 23, pp. 519-535, May 2006.

[63] D. G. Beer, S. L. R. Kardia, C. C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. A. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. G. Taylor, M. D. Iannettoni, M. B. Orringer, and S. Hanash, "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature Medicine,* vol. 8, pp. 816-824, Aug 2002.

[64] E. Huang, S. H. Cheng, H. Dressman, J. Pittman, M. H. Tsou, C. F. Horng, A. Bild, E. S. Iversen, M. Liao, C. M. Chen, M. West, J. R. Nevins, and A. T. Huang, "Gene expression predictors of breast cancer outcomes," *Lancet,* vol. 361, pp. 1590-1596, May 10 2003.

[65] T. A. Pearson and T. A. Manolio, "How to interpret a genome-wide association study," *JAMA,* vol. 299, pp. 1335-44, Mar 19 2008.

[66] S. A. McCarroll, F. G. Kuruvilla, J. M. Korn, S. Cawley, J. Nemesh, A. Wysoker, M. H. Shapero, P. I. de Bakker, J. B. Maller, A. Kirby, A. L. Elliott, M. Parkin, E. Hubbell, T. Webster, R. Mei, J. Veitch, P. J. Collins, R. Handsaker, S. Lincoln, M. Nizzari, J. Blume, K. W. Jones, R. Rava, M. J. Daly, S. B. Gabriel, and D. Altshuler, "Integrated detection and population-genetic analysis of SNPs and copy number variation," *Nat Genet,* vol. 40, pp. 1166-74, Oct 2008.

[67] M. J. Dunning, M. L. Smith, M. E. Ritchie, and S. Tavare, "beadarray: R classes and methods for Illumina bead-based data," *Bioinformatics,* vol. 23, pp. 2183-4, Aug 15 2007.

[68] T. H. Stokes, J. T. Torrance, H. Li, and M. D. Wang, "ArrayWiki: an enabling technology for sharing public microarray data repositories and meta-analyses," *BMC Bioinformatics,* vol. 9 Suppl 6, p. S18, 2008.

[69] O. Ritter, P. Kocab, M. Senger, and D. Wolf, "Igd - Integrated Environment for Genome Data," *Cytogenetics and Cell Genetics,* vol. 66, pp. 15-16, 1994.

[70] R. D. Stevens, A. J. Robinson, and C. A. Goble, "myGrid: personalised bioinformatics on the information grid," *Bioinformatics,* vol. 19 Suppl 1, pp. i302-4, 2003.

[71] K. H. Buetow, "Cyberinfrastructure: empowering a "third way" in biomedical research," *Science,* vol. 308, pp. 821-4, May 6 2005.

[72] A. M. Chinnaiyan, "VISION: OncoMine and caBIG advance cancer bioinformatics," *Scientist,* vol. 19, pp. 22-+, Apr 2005.

[73] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li, "Taverna: a tool for the composition and enactment of bioinformatics workflows," *Bioinformatics,* vol. 20, pp. 3045-3054, Nov 22 2004.

[74] M. Wilkinson, H. Schoof, R. Ernst, and D. Haase, "BioMOBY Successfully Integrates Distributed Heterogeneous Bioinformatics Web Services. The PlaNet Exemplar Case," *Plant Physiol,* vol. 138, pp. 5-17, May 2005.

[75] M. D. Wilkinson and M. Links, "BioMOBY: an open source biological web services proposal," *Brief Bioinform,* vol. 3, pp. 331-41, Dec 2002.

[76] K. A. Karasavvas, R. Baldock, and A. Burger, "A criticality-based framework for task composition in multi-agent bioinformatics integration systems," *Bioinformatics,* vol. 21, pp. 3155-3163, Jul 15 2005.

[77] J. Sroka, G. Kaczor, J. Tyszkiewicz, and A. M. Kierzek, "XQTav: An XQuery processor for Taverna environment," *Bioinformatics,* vol. 22, pp. 1280-1281, May 2006.

[78] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research,* vol. 30, pp. 207-210, Jan 1 2002.

[79] D. A. Hanauer, D. R. Rhodes, C. Sinha-Kumar, and A. M. Chinnaiyan, "Bioinformatics approaches in the study of cancer," *Curr Mol Med,* vol. 7, pp. 133-41, Feb 2007.

[80] D. R. Rhodes, S. Kalyana-Sundaram, V. Mahavisno, R. Varambally, J. Yu, B. B. Briggs, T. R. Barrette, M. J. Anstet, C. Kincead-Beal, P. Kulkarni, S. Varambally, D. Ghosh, and A. M. Chinnaiyan, "Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles," *Neoplasia,* vol. 9, pp. 166-80, Feb 2007.

[81] D. R. Rhodes, J. J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. M. Chinnaiyan, "ONCOMINE: A cancer microarray database and integrated data-mining platform," *Neoplasia,* vol. 6, pp. 1-6, Jan-Feb 2004.

[82] A. Day, M. R. Carlson, J. Dong, B. D. O'Connor, and S. F. Nelson, "Celsius: a community resource for Affymetrix microarray data," *Genome Biol,* vol. 8, p. R112, 2007.

[83] D. J. Duggan, M. Bittner, Y. D. Chen, P. Meltzer, and J. M. Trent, "Expression profiling using cDNA microarrays," *Nat Genet,* vol. 21, pp. 10-14, Jan 1999.

[84] W. A. Baumgartner, K. B. Cohen, L. M. Fox, G. Acquaah-Mensah, and L. Hunter, "Manual curation is not sufficient for annotation of genomic databases," *Bioinformatics,* vol. 23, pp. I41-I48, Jul 1 2007.

[85] M. Hepp, K. Siorpaes, and D. Bachlechner, "Harvesting Wiki consensus - Using wikipedia entries as vocabulary for knowledge management," *Ieee Internet Computing,* vol. 11, pp. 54-65, Sep-Oct 2007.

[86] D. Tapscott and A. D. Williams, *Wikinomics : how mass collaboration changes everything*. New York: Portfolio, 2006.

[87] B. I. Arshinoff, G. Suen, E. M. Just, S. M. Merchant, W. A. Kibbe, R. L. Chisholm, and R. D. Welch, "Xanthusbase: adapting wikipedia principles to a

model organism database," *Nucleic Acids Research,* vol. 35, pp. D422-D426, Jan 2007.

[88] J. Giles, "Key biology databases go wiki," *Nature,* vol. 445, pp. 691-691, Feb 15 2007.

[89] H. Pearson, "Online methods share insider tricks," *Nature,* vol. 441, p. 678, Jun 8 2006.

[90] D. Giustini, "How Web 2.0 is changing medicine - Is a medical wikipedia the next step?," *British Medical Journal,* vol. 333, pp. 1283-1284, Dec 23 2006.

[91] N. Fernandez-Garcia, J. M. Blazquez-del-Toro, J. A. Fisteus, and L. Sanchez-Fernandez, "A semantic web portal for semantic annotation and search," *Knowledge-Based Intelligent Information and Engineering Systems, Pt 3, Proceedings,* vol. 4253, pp. 580-587, 2006.

[92] P. L. Whetzel, H. Parkinson, H. C. Causton, L. J. Fan, J. Fostel, G. Fragoso, L. Game, M. Heiskanen, N. Morrison, P. Rocca-Serra, S. A. Sansone, C. Taylor, J. White, and C. J. Stoeckert, "The MGED Ontology: a resource for semantics-based description of microarray experiments," *Bioinformatics,* vol. 22, pp. 866-873, Apr 1 2006.

[93] M. Hucka, A. Finney, and e. al., "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models," *Bioinformatics,* vol. 19, pp. 524-531, Mar 1 2003.

[94] J. S. Luciano, "PAX of mind for pathway researchers," *Drug Discovery Today,* vol. 10, pp. 937-942, Jul 1 2005.

[95] L. Stromback and P. Lambrix, "Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX," *Bioinformatics,* vol. 21, pp. 4401-7, Dec 15 2005.

[96] S. Oster, S. Langella, S. Hastings, D. Ervin, R. Madduri, J. Phillips, T. Kurc, F. Siebenlist, P. Covitz, K. Shanbhag, I. Foster, and J. Saltz, "caGrid 1.0: An Enterprise Grid Infrastructure for Biomedical Research," *J Am Med Inform Assoc,* Dec 20 2007.

[97] U. Pfeil, P. Zaphiris, and C. S. Ang, "Cultural differences in collaborative authoring of wikipedia," *Journal of Computer-Mediated Communication,* vol. 12, pp. -, Oct 2006.

[98] P. A. Pevzner, "Educating biologists in the 21st century: bioinformatics scientists versus bioinformatics technicians," *Bioinformatics,* vol. 20, pp. 2159-2161, Sep 22 2004.

[99] M. F. Bertoa, J. M. Troya, and A. Vallecillo, "Measuring the usability of software components," *Journal of Systems and Software,* vol. 79, pp. 427-439, Mar 2006.

[100] A. Pettinen, T. Aho, O. P. Smolander, T. Manninen, A. Saarinen, K. L. Taattola, O. Yli-Harja, and M. L. Linne, "Simulation tools for biochemical networks: evaluation of performance and usability," *Bioinformatics,* vol. 21, pp. 357-363, Feb 1 2005.

[101] H. Kitano, A. Funahashi, Y. Matsuoka, and K. Oda, "Using process diagrams for the graphical representation of biological networks," *Nat Biotechnol,* vol. 23, pp. 961-6, Aug 2005.

[102] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, and G. O. Consortium, "Gene Ontology: tool for the unification of biology," *Nature Genetics,* vol. 25, pp. 25-29, May 2000.

[103] "The Gene Ontology (GO) project in 2006," *Nucleic Acids Res,* vol. 34, pp. D322-6, Jan 1 2006.

[104] S. M. Powsner and E. R. Tufte, "Graphical Summary of Patient Status," *Lancet,* vol. 344, pp. 386-389, Aug 6 1994.

[105] J. J. Wang, H. Li, Y. T. Zhu, M. Yousef, M. Nebozhyn, M. Showe, L. Showe, J. H. Xuan, R. Clarke, and Y. Wang, "VISDA: an open-source caBIG (TM) analytical tool for data clustering and beyond," *Bioinformatics,* vol. 23, pp. 2024-2027, Aug 1 2007.

[106] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Research,* vol. 13, pp. 2498-2504, Nov 2003.

[107] M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P. L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker, and G. D. Bader, "Integration of biological networks and gene expression data using Cytoscape," *Nat Protoc,* vol. 2, pp. 2366-82, 2007.

[108] O. Garcia, C. Saveanu, M. Cline, M. Fromont-Racine, A. Jacquier, B. Schwikowski, and T. Aittokallio, "GOlorize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring," *Bioinformatics,* vol. 23, pp. 394-396, Feb 1 2007.

[109] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, "Haploview: analysis and visualization of LD and haplotype maps," *Bioinformatics,* vol. 21, pp. 263-265, Jan 15 2005.

[110] P. Pavlidis and W. S. Noble, "Matrix2png: a utility for visualizing matrix data," *Bioinformatics,* vol. 19, pp. 295-296, Jan 22 2003.

[111] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, "The KEGG resource for deciphering the genome," *Nucleic Acids Res,* vol. 32, pp. D277-80, Jan 1 2004.

[112] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Res,* vol. 27, pp. 29-34, Jan 1 1999.

[113] J. D. Zhang and S. Wiemann, "KEGGgraph: a graph approach to KEGG PATHWAY in R and Bioconductor," *Bioinformatics,* Mar 23 2009.

[114] N. Kono, K. Arakawa, and M. Tomita, "MEGU: pathway mapping web-service based on KEGG and SVG," *In Silico Biol,* vol. 6, pp. 621-5, 2006.

[115] K. Arakawa, N. Kono, Y. Yamada, H. Mori, and M. Tomita, "KEGG-based pathway visualization tool for complex omics data," *In Silico Biol,* vol. 5, pp. 419-23, 2005.

[116] B. R. Zeeberg, W. M. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, and J. N. Weinstein, "GoMiner: a resource for biological interpretation of genomic and proteomic data," *Genome Biology,* vol. 4, p. R28, 2003.

[117] S. W. Doniger, N. Salomonis, K. D. Dahlquist, K. Vranizan, S. C. Lawlor, and B. R. Conklin, "MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data," *Genome Biology,* vol. 4, pp. -, 2003.

[118] N. Salomonis, K. Hanspers, A. C. Zambon, K. Vranizan, S. C. Lawlor, K. D. Dahlquist, S. W. Doniger, J. Stuart, B. R. Conklin, and A. R. Pico, "GenMAPP 2: new features and resources for pathway analysis," *BMC Bioinformatics,* vol. 8, p. 217, 2007.

[119] K. D. Dahlquist, N. Salomonis, K. Vranizan, S. C. Lawlor, and B. R. Conklin, "GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways," *Nat Genet,* vol. 31, pp. 19-20, May 2002.

[120] P. T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W. L. Marks, J. Goncalves, S. Markel, D. Iordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B. J. Aronow, A. Robinson, D. Bassett, C. J. Stoeckert,

Jr., and A. Brazma, "Design and implementation of microarray gene expression markup language (MAGE-ML)," *Genome Biol,* vol. 3, p. RESEARCH0046, Aug 23 2002.

[121] T. F. Rayner, P. Rocca-Serra, P. T. Spellman, H. C. Causton, A. Farne, E. Holloway, R. A. Irizarry, J. M. Liu, D. S. Maier, M. Miller, K. Petersen, J. Quackenbush, G. Sherlock, C. J. Stoeckert, J. White, P. L. Whetzel, F. Wymore, H. Parkinson, U. Sarkans, C. A. Ball, and A. Brazma, "A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB," *BMC Bioinformatics,* vol. 7, pp. -, Nov 6 2006.

[122] H. J. Chung, M. Kim, C. H. Park, J. Kim, and J. H. Kim, "ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics," *Nucleic Acids Research,* vol. 32, pp. W460-W464, Jul 1 2004.

[123] R. Kerkhoven, F. H. J. van Enckevort, J. Boekhorst, D. Molenaar, and R. J. Siezen, "Visualization for genomics: The microbial genome viewer," *Bioinformatics,* vol. 20, pp. 1812-1814, Jul 22 2004.

[124] R. H. Landau, D. Vediner, P. Wattanakasiwich, and K. R. Kyle, "Future scientific digital documents with MathML, XML, and SVG," *Computing in Science & Engineering,* vol. 4, pp. 77-85, Mar-Apr 2002.

[125] Y. Luo and S. Lonardi, "Storage and transmission of microarray images," *Drug Discovery Today,* vol. 10, pp. 1689-1695, Dec 2005.

[126] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi, "The large-scale organization of metabolic networks," *Nature,* vol. 407, pp. 651-654, Oct 5 2000.

[127] R. Albert, H. Jeong, and A. L. Barabasi, "Error and attack tolerance of complex networks," *Nature,* vol. 406, pp. 378-382, Jul 27 2000.

[128] K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabasi, "The human disease network," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 104, pp. 8685-8690, May 22 2007.

[129] G. Mancosu, M. Cosso, F. Marras, C. C. Borlino, G. Ledda, T. Manias, M. Adamo, D. Serra, P. Melis, and M. Pirastu, "Browsing isolated population data," *BMC Bioinformatics,* vol. 6, pp. -, Dec 1 2005.

[130] D. R. Maddison, K. S. Schulz, and W. P. Maddison, "The Tree of Life Web Project," *Zootaxa,* pp. 19-40, 2007.

[131] M. J. Ramirez, J. A. Coddington, W. P. Maddison, P. E. Midford, L. Prendini, J. Miller, C. E. Griswold, G. Hormiga, P. Sierwald, N. Scharff, S. P. Benjamin, and

W. C. Wheeler, "Linking of digital images to phylogenetic data matrices using a morphological ontology," *Systematic Biology,* vol. 56, pp. 283-294, Apr 2007.

[132] E. R. Gansner, E. Koutsofios, S. C. North, and K. P. Vo, "A Technique for Drawing Directed-Graphs," *Ieee Transactions on Software Engineering,* vol. 19, pp. 214-230, Mar 1993.

[133] C. S. Carlson, M. A. Eberle, L. Kruglyak, and D. A. Nickerson, "Mapping complex disease loci in whole-genome association studies," *Nature,* vol. 429, pp. 446-52, May 27 2004.

[134] P. Khatri and S. Draghici, "Ontological analysis of gene expression data: current tools, limitations, and open problems," *Bioinformatics,* vol. 21, pp. 3587-95, Sep 15 2005.

[135] D. B. Searls, "Data integration: challenges for drug discovery," *Nat Rev Drug Discov,* vol. 4, pp. 45-58, Jan 2005.

[136] R. Tibshirani, "A simple method for assessing sample sizes in microarray experiments," *BMC Bioinformatics,* vol. 7, pp. -, Mar 2 2006.

[137] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 98, pp. 5116-5121, Apr 24 2001.

[138] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 102, pp. 15545-15550, Oct 25 2005.

[139] V. K. Mootha, C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop, "PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," *Nat Genet,* vol. 34, pp. 267-273, Jul 2003.

[140] R. A. Fisher, *Statistical methods for research workers*, 12th ed. Edinburgh,: Oliver and Boyd, 1954.

[141] W. M. Feng, G. Tuteja, and M. D. Wang, "EGOMiner: A Genomics and Proteomics Data Computation and Interpretation System for Biomedical Applications," in *IEEE Computational Systems Bioinformatics Conference*, Stanford, CA, 2004.

[142] S. Schuster, D. A. Fell, and T. Dandekar, "A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks," *Nature Biotechnology,* vol. 18, pp. 326-332, Mar 2000.

[143] E. Fahy, S. Subramaniam, H. A. Brown, C. K. Glass, A. H. Merrill, R. C. Murphy, C. R. H. Raetz, D. W. Russell, Y. Seyama, W. Shaw, T. Shimizu, F. Spener, G. van Meer, M. S. VanNieuwenhze, S. H. White, J. L. Witztum, and E. A. Dennis, "A comprehensive classification system for lipids," *Journal of Lipid Research,* vol. 46, pp. 839-861, May 2005.

[144] M. C. Sullards, J. C. Allegood, S. Kelly, E. Wang, C. A. Haynes, H. Park, Y. Chen, and A. H. Merrill, "Structure-specific, quantitative methods for analysis of sphingolipids by liquid chromatography-tandem mass spectrometry: "Inside-Out" sphingolipidomics," *Lipidomics and Bioactive Lipids: Mass-Spectrometry-Based Lipid Analysis,* vol. 432, pp. 83-115, 2007.

[145] A. H. Merrill, M. C. Sullards, J. C. Allegood, S. Kelly, and E. Wang, "Sphingolipidomics: High-throughput, structure-specific, and quantitative analysis of sphingolipids by liquid chromatography tandem mass spectrometry," *Methods,* vol. 36, pp. 207-224, Jun 2005.

[146] A. H. Merrill, M. D. Wang, M. Park, and M. C. Sullards, "(Glyco)sphingolipidology: an amazing challenge and opportunity for systems biology," *Trends in Biochemical Sciences,* vol. 32, pp. 457-468, Oct 2007.

[147] M. Sud, E. Fahy, D. Cotter, A. Brown, E. A. Dennis, C. K. Glass, A. H. Merrill, R. C. Murphy, C. R. H. Raetz, D. W. Russell, and S. Subramaniam, "LMSD: LIPID MAPS structure database," *Nucleic Acids Research,* vol. 35, pp. D527-D532, Jan 2007.

[148] S. Baykoucheva, "A New Era in chemical information: Pubchem, discoverygate, and chemistry central," *Online,* vol. 31, pp. 16-20, Sep-Oct 2007.

[149] A. Funahashi, M. Morohashi, H. Kitano, and N. Tanimura, "CellDesigner: a process diagram editor for gene-regulatory and biochemical networks," *BIOSILICO,* vol. 1, p. 159, 2003.

[150] S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer, "COPASI--a COmplex PAthway SImulator," *Bioinformatics,* vol. 22, pp. 3067-74, Dec 15 2006.

[151] J. B. Bassingthwaighte, "Strategies for the physiome project," *Ann Biomed Eng,* vol. 28, pp. 1043-58, Aug 2000.

[152] P. Hossler, L. T. Goh, M. M. Lee, and W. S. Hu, "GlycoVis: Visualizing glycan distribution in the protein N-glycosylation pathway in mammalian cells," *Biotechnology and Bioengineering,* vol. 95, pp. 946-960, Dec 5 2006.

[153] U. Brandes, T. Dwyer, and F. Schreiber, "Visualizing related metabolic pathways in two and a half dimensions (Long Paper)," *Graph Drawing,* vol. 2912, pp. 111-122, 2004.

[154] J. Lamping and R. Rao, "The hyperbolic browser: A focus plus context technique for visualizing large hierarchies," *Journal of Visual Languages and Computing,* vol. 7, pp. 33-55, Mar 1996.

[155] A. H. Merrill, Jr., T. H. Stokes, A. Momin, H. Park, B. J. Portz, S. Kelly, E. Wang, M. C. Sullards, and M. D. Wang, "Sphingolipidomics: a valuable tool for understanding the roles of sphingolipids in biology and disease," *J. Lipid Res.,* pp. R800073-JLR200, November 21, 2008 2008.

[156] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proc Natl Acad Sci U S A,* vol. 103, pp. 5923-8, Apr 11 2006.

[157] E. Ntzani and J. Ioannidis, "Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment," *Lancet,* vol. 362, pp. 1439-1444, 2003.

[158] E. Marshall, "Getting the noise out of gene arrays," *Science,* vol. 306, pp. 630-1, Oct 22 2004.

[159] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *Lancet,* vol. 365, 2005.

[160] S. Michiels, S. Koscielny, and C. Hill, "Interpretation of microarray data in cancer," *British Journal of Cancer,* vol. 96, pp. 1155-1158, 2007.

[161] J. P. Ioannidis, "Microarrays and molecular research: noise discovery?," *Lancet,* vol. 365, pp. 454-5, Feb 5-11 2005.

[162] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: is there a unique set?," *Bioinformatics,* vol. 21, pp. 171-8, Jan 15 2005.

[163] J. Couzin, "Genomics. Microarray data reproduced, but some concerns remain," *Science,* vol. 313, p. 1559, Sep 15 2006.

[164] M. Eisenstein, "Microarrays: quality control," *Nature,* vol. 442, pp. 1067-70, Aug 31 2006.

[165] K. R. Hess, K. Anderson, W. F. Symmans, V. Valero, N. Ibrahim, J. A. Mejia, D. Booser, R. L. Theriault, A. U. Buzdar, and P. J. Dempsey, "Pharmacogenomic Predictor of Sensitivity to Preoperative Chemotherapy With Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide in Breast Cancer," *Journal of Clinical Oncology,* vol. 24, p. 4236, 2006.

[166] J. D. Shaughnessy, Jr., F. Zhan, B. E. Burington, Y. Huang, S. Colla, I. Hanamura, J. P. Stewart, B. Kordsmeier, C. Randolph, D. R. Williams, Y. Xiao, H. Xu, J. Epstein, E. Anaissie, S. G. Krishna, M. Cottler-Fox, K. Hollmig, A. Mohiuddin, M. Pineda-Roman, G. Tricot, F. van Rhee, J. Sawyer, Y. Alsayed, R. Walker, M. Zangari, J. Crowley, and B. Barlogie, "A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1," *Blood,* vol. 109, pp. 2276-84, Mar 15 2007.

[167] A. Oberthuer, F. Berthold, P. Warnat, B. Hero, Y. Kahlert, R. Spitz, K. Ernestus, R. Konig, S. Haas, R. Eils, M. Schwab, B. Brors, F. Westermann, and M. Fischer, "Customized oligonucleotide microarray gene expression-based classification of neuroblastoma patients outperforms current clinical risk stratification," *J Clin Oncol,* vol. 24, pp. 5070-8, Nov 1 2006.

[168] "The Cancer Biomedical Informatics Grid (caBIG): infrastructure and applications for a worldwide research community," *Stud Health Technol Inform,* vol. 129, pp. 330-4, 2007.

[169] [Anon], "Making data dreams come true," *Nature,* vol. 428, pp. 239-239, Mar 18 2004.

[170] P. A. Covitz, F. Hartel, C. Schaefer, S. De Coronado, G. Fragoso, H. Sahni, S. Gustafson, and K. H. Buetow, "caCORE: A common infrastructure for cancer informatics," *Bioinformatics,* vol. 19, pp. 2404-2412, Dec 12 2003.

[171] W. Sanchez, B. Gilman, M. Kher, S. Lagou, and P. Covitz, "CaGRID White Paper (Cancer Biomedical Informatics Grid Prototype Project)," N. C. I. C. f. Bioinformatics, Ed.: Internet, 2004.

[172] J. Melamed, M. W. Datta, M. J. Becich, J. M. Orenstein, R. Dhir, S. Silver, M. Fidelia-Lambert, A. Kadjacsy-Balla, V. Macias, A. Patel, P. D. Walden, M. C. Bosland, and J. J. Berman, "The cooperative prostate cancer tissue resource: a specimen and data resource for cancer researchers," *Clin Cancer Res,* vol. 10, pp. 4614-21, Jul 15 2004.

[173] B. R. Packer, M. Yeager, L. Burdett, R. Welch, M. Beerman, L. Qi, H. Sicotte, B. Staats, M. Acharya, A. Crenshaw, A. Eckert, V. Puri, D. S. Gerhard, and S. J. Chanock, "SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes," *Nucleic Acids Res,* vol. 34, pp. D617-21, Jan 1 2006.

[174] M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, and J. P. Mesirov, "GenePattern 2.0," *Nat Genet,* vol. 38, pp. 500-1, May 2006.

[175] A. Bouchie, "Coming soon: a global grid for cancer research," *Nat Biotechnol,* vol. 22, pp. 1071-3, Sep 2004.

[176] M. E. Mills, "Linkage of patient records to support continuity of care: Issues and future directions," *Stud Health Technol Inform,* vol. 122, pp. 320-4, 2006.

[177] S. Ruping, S. Sfakianakis, and M. Tsiknakis, "Extending workflow management for knowledge discovery in clinico-genomic data," *Stud Health Technol Inform,* vol. 126, pp. 184-93, 2007.

[178] F. Reddington, A. Ajose-Adeogun, and R. Clark, "The UK National Cancer Research Institute (NCRI) Informatics Initiative: promoting partnership in cancer research," *Hum Mutat,* vol. 28, pp. 1151-5, Dec 2007.

[179] V. Breton, R. Medina, and J. Montagnat, "DataGrid, prototype of a biomedical grid," *Methods Inf Med,* vol. 42, pp. 143-7, 2003.

[180] A. Szalay and J. Gray, "The world-wide telescope," *Science,* vol. 293, pp. 2037-2040, Sep 14 2001.

[181] J. C. Murray, K. H. Buetow, J. L. Weber, S. Ludwigsen, T. Scherpbier-Heddema, F. Manion, J. Quillen, V. C. Sheffield, S. Sunden, G. M. Duyk, and et al., "A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC)," *Science,* vol. 265, pp. 2049-54, Sep 30 1994.

[182] J. Phillips, R. Chilukuri, G. Fragoso, D. Warzel, and P. A. Covitz, "The caCORE Software Development Kit: streamlining construction of interoperable biomedical information services," *BMC Med Inform Decis Mak,* vol. 6, p. 2, 2006.

[183] F. W. Hartel, S. de Coronado, R. Dionne, G. Fragoso, and J. Golbeck, "Modeling a description logic vocabulary for cancer research," *J Biomed Inform,* vol. 38, pp. 114-29, Apr 2005.

[184] "CaBIG Compatibility Guidelines," N. C. I. C. f. Bioinformatics, Ed., 2004.

[185] F. Martin-Sanchez, V. Lopez-Alonso, I. Hermosilla-Gimeno, and G. Lopez-Campos, "A Primer in Knowledge Management for Nanoinformatics in Medicine," in *Lecture Notes in Computer Science*. vol. 5178 Berlin/Heidelberg: Springer, 2008, pp. 66-72.

[186] A. H. Merrill, Jr., T. H. Stokes, A. Momin, H. Park, B. J. Portz, S. Kelly, E. Wang, M. C. Sullards, and M. D. Wang, "Sphingolipidomics: a valuable tool for understanding the roles of sphingolipids in biology and disease," *J Lipid Res,* Nov 21 2008.

VITA

TODD H. STOKES

STOKES was born in Albany, Georgia. He attended public schools in Forsyth, Georgia, and received a B.S. in Computer Engineering from Georgia Institute of Technology, Atlanta, Georgia in 2000. While a student at Georgia Tech, he worked for five semesters at the Georgia Tech Research Institute (GTRI). His technical interests at this time were user interfaces, network programming and distributed computing. After graduation, he worked for four years for the Home Depot, Inc. as a software developer with the title of Senior Systems Engineer. It was at Home Depot that he realized the relatively massive importance of high-quality data over high-quality software interfaces.

In 2004, he returned to Georgia Tech to pursue a doctorate in Bioengineering, inspired by so many experts saying that biomedicine would be the next field to undergo the data revolution. His goal in returning was to develop tools that deliver the same quality, immediacy, and interactivity of information to medical doctors that were being demanded by high-level executives in corporate America. While working as a graduate assistant at the Biomedical Informatics and Bioimaging Laboratory (BioMIBLab), he participated in an NSF fellowship program to encourage commercialization of Georgia Tech research. Along with a team of four members, he competed in the 2008 Georgia Tech Business Plan Competition with a company named omniBioSuite. The omniBioSuite product was a cancer patient status report that could accurately update life expectancy and help patients weigh potential outcomes from various treatment options based on each successive laboratory test result. When he is not working on his research, Mr. Stokes enjoys travel, home improvement, home-brewing, creative writing and studying the business of technology.